179

Vol. 46, No. 3 May, 1949

Psychological Bulletin

EDITED BY

LYLE H. LANIER, NEW YORK UNIVERSITY

CONTENTS

PUBLISHED BI-MONTHLY BY

THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

1515 Massachusetts Ave., N.W., Washington 5, D.C.

Subscription price, $7.00 per year, single issue, $1.25.

Entered as second class mail matter at the post office at Washington, D.C., under the act of March 3, 1879. Additional entry at the post office at Menasha, Wisconsin. Acceptance for mailing at special rate of postage provided for in Section 538, act of February 28, 1925, authorized August 6, 1947.

Vol. 46, No. 3 May, 1949

Psychological Bulletin

PROBLEMS OF PSYCHOMETRIC SCATTER ANALYSIS

JOSEPH JASTAK
Delaware State Hospital

Introduction

Whether tests are administered in omnibus age scales like the Stanford-Binet or in homogeneous point scales like the Wechsler-Bellevue, their successes and failures always exhibit a measurable range of difficulty and a characteristic scattering of responses. The scatter records of individual cases depend partly on the way in which the scale is standardized and partly on the personality problems of the examined person. The possible clinical value of scatter analysis has been hinted at by the earliest experimenters in psychometrics. Its qualitative evaluation has gained momentum in recent years, reaching its culmination point in Rapaport's publication (26) on diagnostic psychological testing and Thurstone's volume (33) on factor analysis.

The main reason for the psychologist's persistent preoccupation with test scatter is the amount of valuable information it yields as a supplement to quantitative indices of brightness. Personality diagnoses often have a basis in the so-called intuitive interpretation of scatter whether clinicians are aware of it or not.

Theoretical Considerations

The clinical usefulness of scatter is in part contingent upon the theoretical viewpoint held by the examiner. On the other hand, the systematic study of scatter may significantly contribute to future changes in psychometric theory (19).

Two general approaches to the problem of response variations on tests are distinguishable. Brody (6), Roe and Shakow (27), Richmond and Kendig (17) and those who standardize tests (32, 34) cling to the idea that psychometric tests measure intelligence. Mental disorders reveal themselves, according to these students of scatter, through their indirect effects on cognitive processes measured by tests. Those engaged in factor analysis (33, 37) hold similar views in the interpretation of group factors. They consider the extracted factors to be organizational

features within the sphere of intelligence. Though non-intellectual personality traits and group factors are being mentioned together with increasing frequency, their relationship is rarely clarified to the satisfaction of the applied psychologist.

Bijou (**5**), Jastak (**13, 14, 15, 16**), Piotrowski (**22**), Rapaport (**26**), and Schafer (**29**) either imply or express the view that intrinsic personality traits may be appraised through test scatter directly without the medium of intellect. Speaking of signs of invalidity, Piotrowski (**22**) finds that some test failures may not be regarded as legitimate symptoms of lack of intelligence. Instead, they are expressions of unorganized or disordered personality functioning in the cognitive, affective, and instinctual areas of behavior. Rapaport (**26**) and Schafer (**29**) point out the dynamic character differences reflected in the relationships between scores of various homogeneous scales. Jastak (**16**) finds that intelligence accounts for only 20 to 25 per cent of the variance of any one test and that the remaining variance must be accounted for by attributes and factors independent of intellectual level.

The differences between these two schools of thought may not appear to be great. Members of both groups accept scatter analysis as a major clinical tool for the understanding of individual differences. However, the personalistic approach probably assumes less and explains more. Every psychometric response has some sort of clinical validity. What it is valid for may not be accurately determined by a priori assumptions but by actual analysis. The traditional pars-pro-toto theory of intelligence artificially limits the experimental horizons of applied psychology. It ignores the Gestalt principle that every human act is the result of the total personality complex. If such traits as introversion, aggression, compulsion, suggestibility and instability affect human adjustments, it is unthinkable that they do not have a direct influence on test scores.

The intelligence theory often falsely identifies cognition with intelligence, confusing an overt act of behavior with a scientific abstraction. It also leads to the juxtaposition of intelligence and emotion which again fails to heed the distinction between theoretical constructs and concrete mechanisms (**8**). Furthermore, it delays the experimental verification of the theory of functional unities or independent traits which are likely to be productive of more fertile ideas in psychometrics than has been the rigid and monodimensional intelligence hypothesis.

Let us assume, for the sake of argument, that behavior is determined by two traits: (1) degree of intelligence, and (2) degree of sanity. If, upon further study, it is found that all degrees of intelligence

occur at a constant level of sanity and that all degrees of sanity occur at a constant level of intelligence (**8**), then the two traits are for all practical purposes independent. If this be true, then intelligence has no direct effect on sanity, and sanity has no direct effect on intelligence. High intelligence does not prevent anyone from becoming or being insane. Insanity does not cause a person to lose his intelligence. Low intelligence may be associated with a high degree of sanity. To borrow a phrase from factor analysis, intelligence and sanity are placed orthogonally to each other in a random sampling of the population. Nevertheless, intelligence and sanity profoundly influence all adjustments simultaneously and differentially. Both may be positively correlated with tests but uncorrelated with each other. It is probable that scatter varies in extent and nature with the combination of the two traits in each individual.

Objective evidence for such a conclusion has so far not been provided. The results of scatter studies are contradictory and confusing. The limitations of tests in the detection of mental disturbances, discussed by Magaret and Wright (**20**), may not be inherent in scatter analysis as such but in the methods and theories employed in its measurement. Psychometric science may successfully catch up with the dictates of common sense if its scope of inquiry is extended without prejudice to all basic qualities of the personality.

The number of traits directly measured by tests are more than two. They are probably different from those used in our simplified example even though these two variables form the central point of interest of nearly all investigators of scatter. The value of mental tests will be greatly enhanced as soon as we stop deciding beforehand what we wish to measure and confine ourselves to the empirical investigation of what is actually being measured.

In the process of developing a qualitative and quantitative system of scatter analysis certain precautions should be taken to obviate extreme swings of the pendulum from a rigid and faulty statistical base to the loose and vague hypothesizing such as is typical of the projective techniques. Scatter analysis may fail to satisfy the requirements of an objective diagnostic aid if certain theoretical pitfalls and technical difficulties are overlooked in its study. The following pages will be devoted to the discussion of the most salient problems associated with scatter analysis.

Scale Standardization and Scatter

Thorndike once said that whatever exists, exists in a certain amount. The numerous attempts at quantifying scattering attest to the ac-

curacy and wisdom of his observation. If scatter is orderly (16, 26, 29) and represents some consistent traits in the examined individual, then such traits should ultimately be measurable through indices of response variability. Despite the discouraging results of scatter analysis on the Stanford-Binet (11, 17, 27), the process is being repeated on an even larger scale with the Wechsler-Bellevue tests.

The Wechsler-Bellevue Scale has several important advantages over the Stanford-Binet Scale. It measures eleven separate abilities in fairly well-graded sub-scales. It has intra-test homogeneity and inter-test heterogeneity. This cannot be said of the Stanford-Binet Scale. In fact, the latter test measures with fair consistency only one type of ability, which is closely related to Thurstone's (33) verbal group factor.

According to McNemar (21), all items of the Stanford Scale are saturated with one general factor. This factor is sufficient to account for most of the inter-correlations between items except at four specified age levels (2, 2½, 6, 18). McNemar also finds a motor factor and a memory factor which may be responsible for some test variability at certain age levels. He warns, however, that the sampling of tests in each of these areas is highly unreliable, and therefore no sweeping pronouncements concerning special defects may be made on the basis of scatter. McNemar's analysis clearly indicates that the Stanford-Binet Scale lacks the much desired inter-test heterogeneity. It would have been a superior diagnostic tool if substantial samplings of many different abilities had been included at all levels.

The high correlations of most items with the general factor may be said to be spurious because of the biased ability sampling of the scale. What McNemar calls a general factor may in reality be a combination of several factors. The differences in common factor saturation between the various items are sufficiently great and consistent to posit the existence of a verbal group factor in addition to the general factor. This hidden group factor, though as constant and stable as the general factor, is responsible for a great deal of scattering on the Stanford-Binet Scale in abnormal cases.

From our experiences with the Stanford-Binet Scale it may be concluded that attempts to assemble test items which are highly correlated with each other and with the whole scale should be abandoned in favor of less correlated but clinically more valuable items. A saturation coefficient of over .60 with any factor should be viewed with considerable suspicion, as it probably represents a statistical distortion produced by improper test samplings or by criterion contamination.

If the theory of multiple traits (16, 33) holds, then high correlations

between tests and factors or traits are little more than standardization artifacts. For the relationships between tests and traits, whether general or specific, are mediocre at best.

Since the centroid verbal factor is so prominent and extensive in the Stanford-Binet Scale, it is likely to misrepresent the mental levels of all intelligent persons who are inferior in this dynamic trait. This is one of the most serious shortcomings of the Scale. The I.Q.'s obtained from such a scale should be used with utmost caution in all diagnostic work with problem children.

High saturation of tests with factors supports the frequently recurring idea that there are pure tests of this and that factor (**8**, **37**). Eysenck, for example, uses the Raven Matrix test as a measure of intelligence because it is supposed to be a test of pure "g." This method of measuring intelligence contradicts Eysenck's very poignant argument that each test is the result of at least four factors, one general, one centroid, one specific, and one error factor. The theory of manifold causation is so reasonable and pertinent that it should relegate the idea of pure factorial tests to the realm of statistical myths. The reviewer suspects that the Raven Matrix test measures several important group factors in addition to the general factor. To find that the Matrix test is a relatively pure test of "g" must surely be due to faulty experimental design. The same may be said of tests which are supposed to be pure measures of some group factor (**37**). The existence of pure tests of either general or group factors cannot be reconciled with psychological realities in clinical practice.

One of the greatest, and as yet unsolved, problems of the experimental laboratory and clinic alike is the coordinate interpretation of the results from many different tests. Rapaport (**26**) and Eysenck (**8**) treat the large number of different tests and techniques without making direct comparisons between them. This procedural atomism not only breaks up the diagnostic study into artificial and overlapping sections but also retards the final determination of what the various tests measure and how well they measure what they do. When the Bellevue and Babcock tests (**26**) differentiate between neurotics and schizophrenics, they may not necessarily do it in the same manner. The factors involved may not be identical or even similar. When tests of vocabulary, social history, persistence, choline esterase, and body build differentiate between Eysenck's (**8**) hysterics and dysthymics, it does not follow that all these methods are good and valid tests of the same temperamental factor as Eysenck seems to imply. The way to find out what is being measured by all these tests is to administer them to the

same experimental population and to a control group and then subject them to an over-all factor analysis. In this manner, the similarity or dissimilarity of the analytic criteria would soon become apparent. It would also facilitate the discovery of several differential signs based not on the nature of the testing procedure but on the nature of the psychic traits or genotypes measured.

Fourfold Analysis of Tests

Psychometric batteries composed of a large variety of independently scaled, homogeneous abilities yielding separate scores are conducive to a fourfold analysis of test variability.

First, they permit the comparison of one individual in any ability with the norm of many other persons of the same age in that ability. This inter-personal index is similar to the I.Q. except that reliable scores obtained from several discrete test functions increase the analytic possibilities. They provide a multidimensional system of comparison.

Second, homogeneous subscales measure the intra-personal organization of functions which are directly and intimately related to dynamic personality traits. The discrepancies between scores within each individual record are of great diagnostic importance (**4**, **5**, **8**, **16**, **17**, **23**, **26**, **28**, **29**, **30**, **31**, **34**).

Third, the response variations of a person within each subscale may be used in the interpretation of test results. Two persons obtaining identical results on an information test, for example, may have arrived at this score in a totally different manner. Indeed, identical score patterns within a scale are a rare exception in records with identical final achievements. Eysenck (**8**) attempted to measure intra-test scatter in relation to the Matrix test. Because of the method he used and because of the lack of direct inter-test comparisons his results were negative as far as differentiation between his diagnostic groups was concerned. The reviewer (**16**) believes that intra-test scatter may contribute to the understanding of personality dynamics. It is to some extent related to inter-test discrepancies. It defines the nature of inter-test variability.

Fourth, a test response may be analyzed in regard to several aspects which have so far not received sufficient objective evaluation. They are, however, given some attention in Rapaport's (**26**) and Schafer's (**29**) writings. Reaction times, rate of production, amount of output, contents of response, form of response, nature and number of errors preceding the solution, and projective qualities deriving from emotionally tinged, personal experiences are important factors in all tests. These facets of a response undoubtedly influence the clinician's judgment even though they receive no methodical statistical analysis. That such an analysis should ultimately be made to keep diagnostic inferences within reasonable bounds is self-evident. It is precisely these unobjectified and intangible qualities of a response that lend themselves to the making of far-fetched and exaggerated claims. Clinical rationalizations may be kept to a minimum through the adoption of multiple scoring techniques which, if necessary, overlap some or all of the subscales of a battery. The appraisal of characteristic and psychologically uniform response deviations appearing intermit-

tently throughout the battery may lead to the isolation of a number of helpful diagnostic signs.

Like life itself, tests are neither fully structured nor unstructured. Projective elements are liable to occur on all tests. A vocabulary test, properly standardized, may be equally as projective as the Rorschach test (**18**). Furthermore, the measurement of projective qualities by means of vocabulary and other psychometric tests is likely to be more precise and more meaningful than are similar measurements from so-called unstructured tests. A change in the theoretical foundation of the Binet type of test will produce more satisfactory diagnostic and therapeutic indicators than does its replacement by tests which elude scientific treatment.

The Wechsler-Bellevue Scale (**34**) or any similarly constructed battery (**14, 15**) may be a valuable tool of multiple response analysis if the necessary precautions are applied to the theoretical and technical problems of scatter measurement. The potentialities inherent in such an analysis are great and numerous. Variability studies are certainly as important as the final test scores and more informative than intelligence quotients.

A good psychometric scale should, indeed, be standardized for the various types of scatter in addition to providing achievement norms. Before this is done, psychologists will have to be satisfied with cluster analysis as an after-thought rather than as a primary diagnostic aid.

External Criteria of Validation

Scatter is more often observed in mentally ill or disordered persons than in normals. Since most scatter studies come from psychiatric clinics and mental hospitals, it is natural to find them evaluated in terms of psychiatric disease entities. Two objections may be raised to such a procedure. First, attendance at a clinic or commitment to a hospital does not make a person abnormal and, vice versa, not being a psychiatric patient does not make a person normal. Second, psychiatric classifications of mental illness are mono-dimensional and therefore do not fully agree with all the facts in the given case. Psychiatrists generally consider their present-day nosology outdated and unsatisfactory. A great deal of research will have to be done to objectify the syndromes associated with mental illness and personality disorganization. The very existence of disease unities such as are postulated in psychiatric diagnoses awaits experimental verification.

The criteria differentiating one illness from another are vague and fluid. The diagnosis therefore depends largely on the school and sub-

jective bias of the individual psychiatrist. Some students of scatter analysis (**8**, **16**, **26**, **29**, **34**) are fully aware of the implications inherent in the use of comparative standards of low or dubious validity. There may be substantial agreement between psychometric findings and psychiatric evaluations outside the reference frame of official nomenclatures. However, most scatter analyses are based directly on official diagnostic categories. The outcome of such studies hinges as much on the objectivity of the psychiatric grouping as it does on psychometric analysis.

This dilemma may be overcome either by abandoning the practice of validating scatter against psychiatric diagnoses as now conceived and employed, or by obtaining from the psychiatrist a comprehensive evaluation of the adjustment pattern of each patient. Another method of improving validation criteria is to give greater weight to the intrinsic relationships of test scores obtained from the experiment. Chemists, physicists, and astronomers depend almost exclusively on criteria of internal consistency set up as a result of well-planned laboratory procedures. Objective psychology has reached a stage at which criteria of internal consistency may be of considerable service. For example, the discrepancy between drawing and arithmetic scores is a more valid criterion of mental adjustment than any extraneous criterion except the actual life developments of the individual case. These developments should be checked and studied by the experimenter directly without complicating the issue by seeking opinions and impressions of inferior value. There may, in fact, be a closer agreement between psychometric test criteria and living realities than is generally assumed to be the case. Psychologists will make their greatest contribution to psychiatry when they emancipate themselves from current psychiatric systems of diagnoses and evolve a rigorous and multiple system of comparison between integrated and unintegrated personalities.

Reference Points

The phenomenon of response scatter is a universal one. It has many degrees and many qualitative variations. It occurs in persons of all ages and types. The general assumption is that scatter is greater in those with abnormal than in those with normal personalities. This assumption will be found to be correct when tests are standardized for scatter on the basis of criteria of internal consistency. One of the simplest and most direct measures of scatter is the range ratio. It is obtained by dividing the lowest score from a large number of subscales by the highest score from the same tests (**16**). The range ratio of the

Wechsler-Bellevue test is considerably smaller in all abnormal groups than it is in the normal population. It indicates that disturbed and unorganized people are more variable, less dependable, and less compact in most of their adjustments (16). The range ratio is an objective measure of the degree of mental fluctuations and general variability. It does not distinguish between different types of scatter within each record.

When attempts are made to determine the scatter clusters between or within tests, the choice of a reference point from which scatter is measured becomes important. The psychological and statistical soundness of this reference point governs the soundness of the final scatter results. It must have statistical stability and psychological relevance. In other words, it must be clinically valid and statistically reliable. Its choice is frequently a function of the personality theory to which the experimenter subscribes.

Three types of reference points have so far been used in the study of test scattering:

1. Deviations of test scores from the stable score of a test which is usually believed to be a good test of original intelligence. Babcock's vocabulary scatter (1, 2) and Eysenck's Matrix test scatter (8) are two examples of this type.
2. Deviations of scores from the mean of all tests or of a group of tests. This is one of the most popular methods of scatter analysis. It has been used by Wechsler (34), Rapaport (26), Schafer (29), and many others.
3. Factor or cluster analysis in which the mean of one cluster is compared with the means of other clusters of tests. Thustone's (33) group factors and Jastak's (16) altitude scatter belong in this category.

Let us scrutinize the practical value of these three procedures in the light of clinical experience and experimental fact.

Vocabulary Scatter

The advantages of comparing vocabulary scores with other test scores are two. The vocabulary yields one of the most stable rating used in psychometric examinations. It is also relatively invulnerable to the presence of disorders and disturbances such as are observed in psychotic persons. This refractoriness of the vocabulary score has been noted very soon after the introduction of tests as a method of measuring intelligence. Wells (35) was the first one to use it in a systematic manner for purposes of measuring the variability of mental patients. Babcock (3, 4) has elaborated the relationship between vocabulary and other tests into her well-known "level-efficiency" theory. The vocabulary test, being uninfluenced by mental disorganization, serves as an index of native capacity, while the highly sensitive efficiency tests, being reduced by mental disturbances, serve as measures of sanity. The rela-

tionships between the two types of tests may thus be employed to determine the degree of deterioration. Babcock (**1**, **2**) has provided experimental evidence for the partial correctness of her theory by testing it on patients suffering from general paresis and dementia praecox.

More recently, Rapaport (**26**) confirmed the diagnostic value of vocabulary scatter by using Babcock's tests in clinical examinations. He states "These statistical data then substantiate the contention that Vocabulary is the least variable and one of the best retained of the subtests; thus it might well serve as an indicator of the original intelligence level, and as a standard of comparison for the other subtests." The two premises underlying this method of scatter analysis are that vocabulary is a good test of intelligence and that its score is insignificantly depressed by mental illness. Though vocabulary has been shown to decline somewhat with age (**34**) and to decrease significantly in cases of organic deterioration (**6**, **7**), its relative invulnerability has been sufficiently documented by actual test throughout the history of clinical psychology. The stability postulate of vocabulary scatter might, with some reservations, be clinically acceptable.

The assumption that the vocabulary is a superior test of intelligence depends on the definition of intelligence and on actual correlations between intelligence and vocabulary. Since a general theory of intelligence has so far not been evolved, the postulate that vocabulary gives valid measures of intelligence remains experimentally unverified. Jastak (**16**) finds that the vocabulary is no better a test of native capacity than is any other well-standardized test. If this is true, then vocabulary scatter should prove quite unsatisfactory and even misleading in a large number of mental patients. It is imperative that the reliability of a score not be confused with its validity. The Wide Range Reading Test (**13**) is even more stable and less vulnerable in mental examinations than is the vocabulary. Yet we would not dare suggest that oral reading is a valid test of pre-psychotic intelligence because no test can be credited with such unusual powers, no matter how stable its score may be. This criticism applies in even greater degree to Raven's Matrix test used by Eysenck (**8**) as a measure of intelligence.

If a person has a defective vocabulary despite good intelligence, then the difference between vocabulary and efficiency tests may be positive or may indicate a favorable degree of personality organization even though the person is severely disorganized. Furthermore, an individual may have an efficiency score of 120 and a vocabulary score of 75 before he becomes psychotic. During psychosis the efficiency score drops 40 points (from 120 to 80), whereas his vocabulary score remains

75. In such cases (not at all rare in clinical practice) the vocabulary scatter is useless as a measure of efficiency for the simple reason that it is not an accurate measure of original intelligence. It is incapable of revealing the decline in efficiency which has actually occurred.

Vocabulary scatter is valuable only in a small number of cases such as are seen in private clinics and other institutions catering to a highly select group of patients. It fails to distinguish between integrated and unintegrated behavior in the majority of cases treated in clinics and hospitals which draw their patients from all segments of the community. The reviewer (**14**) has found that the magnitude and the distribution of vocabulary scores among mental patients are similar to those of normal persons. This would indicate that vocabulary is indeed uninfluenced by mental disturbances. This does not, however, entitle us to conclude that the vocabulary test always furnishes an accurate classification of original intelligence. Mental levels based on vocabulary scores are accurate and valid in not more than 10 per cent of mental patients. Clinicians will, therefore, do well to use the vocabulary test with utmost caution in the determination of intellectual level and also in the determination of efficiency by means of vocabulary scatter.

Babcock (**3**, **4**) has advanced psychometric theory by adding the efficiency dimension to the intelligence quotient. This extension of the usefulness of tests is valuable but it does not go far enough. The human personality is multi-dimensional. The most complete system of psychometric interpretation must therefore be manifold. As Rapaport (**26**) puts it " . . . Vocabulary Scatter . . . is of little value in gauging the general spread of scatter over the scattergram. Our approach to scatter analysis is to consider the relation of all subtests scores to each other. Considering scores in relation to the vocabulary level is only one aspect of this approach." A multi-dimensional approach is not limited to one rigid reference point, but employs many reference points on the basis of their empirical value in test analysis. The recognition of this important tenet will obviate the elevation of one test to the pedestal of an exclusive reference point (**2**, **8**) since it limits the possibilities of meaningful scatter analysis through the study of many other relationships.

Mean Scatter

In appraising the diagnostic and clinical features of the Bellevue Scale, Wechsler (**34**) has found that the deviations of the eleven subtest scores from the means of all scores bear certain consistent relationships to the disorders of psychiatric patients. Wechsler uses differences of individual test scores from the means. Rapaport (**26**) and Schafer

(28, 29, 30) employ a similar method of measuring scatter. This technique is, of course, more subtle and more versatile than vocabulary scatter. It comes close to a multi-dimensional system and satisfies to some extent the criterion of psychological polivalence of test scores. It should prove serviceable if the mean of all scores meets the validity requirements of analysis. The mean of all scores is undoubtedly a very stable reference point. Whether it is psychologically relevant is far less certain. Its relative constancy is an asset, its psychological ambiguity is a liability. The mean of all scores often acts like an unwanted middleman. It cuts the real discrepancies and disregards the direct relationships between tests. The difference of 9 points between the weighted score of 12 on the vocabulary test and the weighted score of 3 on the arithmetic test may be more meaningful than are the two scores of +4 and −5 from a mean of 8 points.

The difficulties of mean scattering increase when the mean is close to the extremes of the distribution of individual weights. In such cases, it is a highly contaminated statistic. It neutralizes the very deviations which are of diagnostic importance. An acutely psychotic or seriously disorganized individual may have a low mean score as a result of his personality defects. The actual discrepancies from the mean, whether cumulated or not, may give no evidence of a disturbance when in reality it is very severe. The more generalized the mental disorder the less efficiently does the mean differentiate between normal and abnormal functioning. The general thesis may be adopted that the use of any reference point for the measurement of a trait is impractical if the trait in question has a variable effect on that reference point. If the mean of all tests, like the vocabulary, were unaffected by personality impairment, it could well serve as a reference axis of scattergrams. Since this requirement is never satisfied in clinical studies, the mean of many tests may not be regarded as a psychologically adequate pivot about which scatter varies in a consistent manner. Like the vocabulary test, it has a great deal of reliability but little validity for the accurate appraisal of scatter. The vocabulary score is relatively uninfluenced by mental disorganization, but it is not a valid measure of original intelligence. The mean score is an invalid index of original intelligence because it is influenced by mental illness.

A third flaw of the mean is that the various types of mental conditions have different effects on it in different persons. It may thus give an accurate picture of a trait in one individual, but obscure the very same trait in another individual.

Cluster Analysis

It is infrequently realized that clinicians and factor analysts are doing the same thing in different ways. When Babcock (**3**) decided to divide her efficiency battery into motor, learning, and memory tests, she was engaging in intuitive factor analysis. When Wechsler (**34**) grouped the subtests of the Bellevue Scale into verbal and performance parts, he assumed that the two types of tests measured different functions. When Rabin (**25**) divided the scores of one cluster of tests by the scores of another cluster to differentiate between schizophrenia and manic-depressive psychosis, he applied a clinical method of factor analysis. When Rapaport (**26**) eliminated the digit span and arithmetic tests from his verbal scatter measures, he employed a legitimate method of cluster analysis which could have been readily confirmed by Thurstone's centroid method of factor analysis.

Factor analysis (**8**, **16**, **33**, **37**) accomplishes by statistical means in a relatively short time what the clinician discovers after many years of hard-won experience. Even then, the clinicians' hunches may be incomplete and unrefined in comparison with the differentiations supplied by factor analysis. That is why Thurstone's group factors may be considered another form of scatter analysis.

Factor analysis has the advantage over other systems in that it reduces the almost infinite complexity of test interrelationships to a smaller number of psychologically pertinent unities. It directs our attention to the weighty fact that scatter analysis has as many facets as there are factors. Scatter may be large in some tests and small in others depending on the personality structure of the examined individual. Unless we learn to differentiate between types of scatter within the same test battery, we are likely to end up with few or no positive signs despite Herculean efforts. When scatter scores are averaged without regard to the clusters in which they occur, their diagnostic significance may disappear in the process of analysis.

The strong points of factor analysis may be summarized as follows:

1. It is an objective and efficient means of grouping psychologically homogeneous tests.
2. It provides a highly differentiated system of scatter analysis within the isolated clusters.
3. It supplies several stable and important reference points that are clinically meaningful.
4. It shows that one sub-test may belong to two or more clusters and that such tests should be used in the appraisal of more than one trait. This latter possibility is especially difficult to gauge by mere clinical impressions.

The clinical application of factor analysis has been limited in the past because of several shortcomings in the interpretation of factors.

1. The naming of factors after test contents does not satisfy the clinical psychologist who thinks more in terms of personality dynamics than in terms of test materials.

2. Factor analysis has not been able to provide factorial scores which are actually uncorrelated with each other. For example, the number factor and the verbal factor always yield scores that are positively correlated. The independence of these factors has not been empirically demonstrated up to this point.

3. Those who believe in group factors ignore the possible existence of one or more general factors which are responsible for the positive correlations found between group factor scores. Unless these general factors are accounted for and held constant, scatter analysis by factoring will be of little avail to the clinician. In fact, the general factor, properly measured, might be used as a stable and valid reference point against which the group factors are evaluated. Such a method was recently suggested by the reviewer (**16**).

4. The scope of factorial interpretations must be enlarged to include the entire repertory of personality traits. Groups factor may in reality represent personality or character traits which are independent of intellectual capacity. Factor analysis is more adequate in its statistical aspects than in matters of mental analysis. To improve its practical usefulness, the following methodological improvements are proposed: (a) the isolation and measurement of the general factor of intelligence; (b) the isolation and identification of group factors with non-intellectual character dynamics; (c) the study of the relationships of the general factor to group factors; (d) The empirical demonstration that the deviations of group factor scores from intelligence are independent of the degree of intelligence; and (e) the empirical demonstration that the deviations of the various group factor scores are independent of each other.

Weighted Scores as Measures of Scatter

Scatter studies of the Wechsler-Bellevue Scale (**34**) are numerous and more objective than are those of any other test scale. However, most of the published studies suffer from a number of technical deficiencies which should here be pointed out. One of these is the almost universal use of weighted scores.

As is well-known, the raw score of each one of the 11 Bellevue subtests is converted into a weighted score to equate the achievements of different persons on the same test. The weighted score is in turn transmuted into an intelligence quotient. The weighted score and the intelligence quotient are statistically equivalent for they are both standard scores. Psychologically, they differ significantly because the weighted score disregards age whereas the quotient does not. The Wechsler quotient is thus a convenient method of keeping age relatively constant. Scatter patterns are more accurate and more meaningful

when expressed in terms of quotients rather than in terms of weighted scores.

Rapaport's (26) 67 schizophrenics have a mean age of 31 years; his 33 depressed patients have a mean age of 49 years. The depressives are on the average 18 years older than are the schizophrenics. Rapaport finds that "No schizophrenic group, not even the Deteriorated Schizophrenics, show such a consistently great impairment of all Performance subtests as the two Depressive Psychosis groups. This scatter pattern is of crucial diagnostic significance for differentiating Depressive Psychoses from Schizophrenia."

If age is kept constant by the use of test quotients, his depressive groups fail to show the consistently great impairment of performance subtests. When all the precautions discussed in this paper are applied to Rapaport's cases, the average performance scatter score (deviation ratio from capacity) for his 44 most chronic and deteriorated schizophrenics is 82.9. The average deviation ratio for his 33 depressives is 82.2. The difference is insignificant. Both scores are significantly below the average (normal) scatter score of 87.4. Rapaport's deteriorated unclassified schizophrenics (N=7) have a scatter score of 74.7 which is below that (78) of his most disorganized depressives (N=8). Rapaport's crucial scatter differences are more a function of the age differences between the diagnostic groups than they are of the groups themselves. Foster (9) has shown that when weighted scores are corrected for age, the originally diagnostic scattergram turns out to be an artifact. The test discrepancies either become smaller or disappear when age is kept constant.

It may be added that the performance deviation ratios of Rapaport's groups are generally high and the discrepancies rather small. They approach the norm of the population at large much more closely than do the ratios of hospital patients studied by the reviewer. This may be explained in two ways. Either Rapaport's cases were less disorganized than are psychiatric patients generally. Or, the diminished scatter is due to the fact that his patients were of high average and superior intelligence. High capacity as such is not a deterrent in the study of scatter. However, our experience indicates that the Bellevue Scale is least satisfactory at the extremes of its ability range. It does not give the superior person sufficient opportunity to demonstrate his true superiority. Nor does it give the stupid person a full chance to show his real level of stupidity. This telescoping effect of the Bellevue Scale is a distinct weakness in all investigations of scatter. Patients are prevented from scattering because of the limitations of the scale at both ends.

Any scatter analysis of the tests of the Bellevue Scale by means of weighted scores is of dubious value, unless the experimental or clinical groups are homogeneous in age or unless test quotients are used to allow for age differences between members of such groups.

Sex Differences

It was seen that age differences may produce typical scattergrams. These patterns are of little value in diagnosis if not corrected for age. Another factor which is likely to disturb the clinical significance of subtest analysis is sex. Eysenck's study of test variability (8) is one of the most satisfactory of all those published in that he always separates the two sexes.

As long as psychometric ratings are expressed in averages from many different abilities, the differences between males and females are neutralized. This is especially true if the battery consists of an equal number of tests favoring males and females. As soon as the averages are broken up into subcomponents, significant sex differences may appear in some of the clusters. Scatter analysis will then show disorganization where it does not exist. Or, it may show normal organization in disturbed individuals. Sex differences may not be ignored in the interpretation of various types of scatter. On the Bellevue test, males tend to be better in information, picture completion, object assembly, arithmetic, and digit span. Females tend to be better in vocabulary, symbol substitution, comprehension, picture arrangement, and block designs.

Vocabulary scatter may, in the light of these differences, mean something different in case of women than in the case of men. If a predominantly female cluster of tests happens to be used in the computation of scatter scores, it may give a distorted picture of personality functioning just because the patient is a man or a woman.

Future psychometric scales, if standardized for scatter, will have to be equated for sex differences.

Arithmetical Differences as Measures of Scatter

It is common practice to express test discrepancies in terms of differences between scores. Thus vocabulary scatter (1, 2, 26) is usually measured by subtracting the vocabulary score from the efficiency score. If the difference is positive or small, the subject is assumed to be normal. If the difference is a large negative one, the patient is thought to be deteriorated. Weighted and sigma scores are used in the calculation of differences, presumably to render them comparable. In the mean scatter method, the subtest score is subtracted from the mean. Babcock (1, 2),

Wechsler (**34**), and Rapaport (**26**) use the method of differences between weighted or standard scores. Rapaport (**26**) mentions the necessity of dealing with positive and negative numbers as bothersome, but this is not the real difficulty of the method.

Wechsler (**34**) points out correctly that identical differences have different meanings at different achievement levels. Thus a difference of 3 points at the weighted score level of 3 to 6 is twice as great as the difference of 3 points at the 9 to 12 level. If this important fact is ignored, the findings of scatter analysis from the method of differences are incomparable. It must be stressed that differences in terms of sigma scores do nothing to alleviate this problem.

Differences are positively correlated with the magnitude of the subtrahend. Wechsler (**34**) overcomes this difficulty by dividing the total weighted scores of ten tests by 40. The quotient thus obtained indicates the significance of the differences between scores at different test levels. When the sum of the weighted scores is 120, a difference of 3 points (120/40) is significant. When the sum of the 10 weighted scores is 60, a difference of only 1½ points (60/40) is significant.

Jastak (**14, 15, 16**) has achieved the same results by dividing directly the smaller standard score by the larger. One of the characteristics of ratios is that they are uncorrelated with the denominator used in their calculation. The best known example illustrating this principle is the intelligence quotient. The I.Q. is by definition uncorrelated with the chronological age. If an index of brightness had been proposed in terms of the plus-minus differences between mental and chronological age, it would have the same unwieldy and misleading properties typical of most scatter measures now in use.

The sums of differences between scores for two persons are not comparable unless both persons have identical subtrahends. The mean-difference scatter of deteriorated schizophrenics is not comparable with that of depressives unless both groups have identical means.

Ratios are generally more useful in the study of the relationships between tests than are differences (**10**). They may, for example, be conveniently employed in measuring the correlational residuals of group factors when the general factor is kept constant (**14, 15, 16**). Dividing the group factor scores by the general factor scores will accomplish this end. The deviation ratios from capacity of each subtest of a homogeneous cluster are positively correlated with each other. On the other hand, the size of the deviation ratios of a cluster from capacity is uncorrelated with capacity. Furthermore, the deviation ratios of one cluster may be shown to be uncorrelated with the deviation ratios of another

cluster. When the personality trait representing a group factor stands in a compensatory relation to intelligence, as happens in some highly selected vocational groups, the correlation between deviation ratios of one cluster and intelligence (or another cluster) may actually be negative. In random samplings, the correlations between general and group factors approach zero.

That human adjustments are governed by compensatory relationships between dynamic traits is more than a popular belief. Scatter and cluster analysis may become appropriate vehicles for demonstrating that beyond the correlational realm of measured abilities there exist traits or factors which are either uncorrelated or which, under certain conditions, may take up opposite poles in relation to over-all personality functioning.

The scientific proof of trait compensation devolves upon one important contingency. Select groups may not be treated as self-sufficient units, but should be compared with the norms of a broader universe—the total population. The experimental controls must be random in regard to all personality traits which characterize people because all these traits share in the production of scatter on tests. The original sampling must be random at all age levels and for both sexes.

Psychosis, Cooperation, and Scatter

Eysenck (8), Rapaport (26), Jastak and Vik (15) speculate about the possible relationship between psychosis and psychometric evidence of disorganization. The consensus of opinion seems to be that psychometric disorganization is more basic and more far-reaching than is the fact of acute psychosis. Abnormal scattergrams may precede the onset of psychosis by many years and may continue unchanged following remission. Many psychiatric remissions are surface phenomena unaccompanied by deep-seated changes in personality organization. The personality deviations measured by tests may serve as the background for the chronic or intermittent exacerbations in the form of delusions, hallucinations, and acute psychotic states.

The correlation between degrees and types of scatter and psychosis is not high, as Wittman (38) and Brody (6) point out. There is no valid reason for assuming that such correlation should be greater than moderate at the very best. Attempts to prove that psychometric disorganization and psychosis are highly correlated is a futile undertaking. Whenever high correlations between psychosis and psychometric scatter are reported, they may be considered the result of ill-designed experiments. A psychosis is not infrequently an acute and temporary condition super-

imposed upon a relatively normal personality. To measure scatter and disorganization in psychotic and deviant persons, the scatter of normal people must be known. The judicious use of internal criteria of consistency will demonstrate that those usually considered normal may not necessarily be so, and obversely, those considered abnormal may not be so either.

Wittman (**36**) expresses the view that psychometric scatter as measured by Babcock's method may be a function of test cooperation on the part of the subject. She is seconded in this view by Brody (**6**). Acceptance of this opinion depends on the definition of cooperation. Inability to cooperate is usually a part symptom of the mental disturbance and therefore need not cause serious concern as a major factor in scattering. Great skill and patience are required to administer tests to disturbed individuals. The results of such tests are fully as valid as those obtained from entirely cooperative persons. When the patient is unwilling to cooperate, the test is incomplete and scatter analysis impossible. There is little else to report except the fact of refusal and the circumstances under which it occurs. The reviewer is inclined to the view that the degree of cooperation is a very minor factor in the production of test scatter.

Psychometric results are invalid as measures of basic personality traits only in cases of profound stupor, seriously reduced consciousness, and complete mental confusion.

Summary

Scatter analysis is a central problem in all psychometric examinations. Objective comparisons may be inter-personal and intra-personal. Inter-individual norms delimit the standing of a person in any ability or trait in comparison with other persons in that ability or trait. Intra-individual norms may be established by the study of inter-test discrepancies and intra-test response patterns. Inter-test discrepancies require a systematic survey of the relationships between many tests and abilities within the same individual. The method of choice is factor analysis complemented by clinical insight and a personalistic approach.

The reference point of scatter analysis should be provided by a psychologically homogeneous and statistically stable trait. A general factor, accurately measured and operationally defined, is the most suitable for this purpose. Neither vocabulary scatter nor mean scatter, though they constitute invulnerable and stable statistics, is clinically applicable to the study of personality traits. The vocabulary test may not be considered an accurate and valid measure of pre-psychotic intelligence.

The mean of all scores, on the other hand, is often a highly contaminated value which reduces the sought-after variations in a variable manner.

Psychometric scales should be standardized for scatter as well as for achievement scores in different abilities. Standardization samplings should be random in regard to all measurable personality traits. External criteria of validation should either be improved or supplanted by criteria of internal consistency determined in the course of the experiment.

The use of weighted scores or similar measures uncorrected for age and sex interferes with the objective appraisal of clinically valuable scatter records. The method of differences between scores and a reference point should be replaced by the method of ratios of one score to the other. The method of ratios eliminates positive correlations between the magnitude of the reference point and the size of the discrepancies from it. It makes the deviations comparable at all levels of achievement.

The most far-reaching change in regard to scatter analysis is the extension of the value of psychometric tests to the field of general personality measurement. The restrictive pars-pro-toto theory of intelligence should be abandoned. Furthermore, the idea that tests of pure factors can be assembled is contrary to the principles of personality unity and simultaneity. It may be safely assumed that any ability whatever is the function of at least four types of factors: general, centroid, specific, and accidental.

BIBLIOGRAPHY

1. Babcock, Harriet. An experiment in the measurement of mental deterioration. *Arch. Psychol. N. Y.*, 1930, No. 117.
2. Babcock, Harriet. *Dementia praecox: a psychological study.* Lancaster, Pa.: Science Press, 1933.
3. Babcock, Harriet. Personality and efficiency of mental functioning. *Amer. J. Orthopsychiat.*, 1940, **10**, 527–531.
4. Babcock, Harriet. The level-efficiency theory of intelligence. *J. Psychol.*, 1941, **11**, 261–270.
5. Bijou, S. W. The psychometric pattern approach as an aid to clinical analysis—a review. *Amer. J. ment. Def.*, 1942, **46**, 354–362.
6. Brody, M. B. A survey of the results of intelligence tests in psychosis. *Brit. J. med. Psychol.*, 1942, **19**, 225–261.
7. Davidson, M. Studies in the application of mental tests to psychotic patients. *Brit. J. med. Psychol.*, 1939, **18**, 44–52.
8. Eysenck, H. J. *Dimensions of personality.* London: Kegan Paul, 1948.
9. Foster, A. Age and the Wechsler-Bellevue scattergraph. *J. clin. Psychol.*, 1947, **3**, 396–397.
10. Hamlin, R., & Abel, T. M. Test patterns of mental defectives skilled in

weaving. *J. appl. Psychol.*, 1938, **22**, 385–389.

11. Harris, A. J., & Shakow, D. The clinical significance of numerical measures of scatter on the Stanford-Binet. *Psychol. Bull.*, 1937, **34**, 134–150.
12. Jastak, J. *Variability of psychometric performances in mental diagnosis.* New York: Privately published, 1934.
13. Jastak, J. School test patterns of clinic children. *Del. State med. J.*, 1938, **10**, 108–111.
14. Jastak, J., & Gordon, G. J. Verbal polarity in dementia precox. *Del. State med. J.*, 1946, **18**, 131–136.
15. Jastak, J., & Vik, E. S. The objective measurement of reality perceptions in dementia precox. *Del. State med. J.*, 1947, **19**, 98–102.
16. Jastak, J. A plan for the objective measurement of character. *J. clin. Psychol.*, 1948, **4**, 170–178.
17. Kendig, Isabelle, & Richmond, Winifred V. *Psychological studies in dementia praecox.* Ann Arbor, Mich.: Edwards Bros., 1940.
18. Klopfer, B., & Kelley, D. M. *The Rorschach technique.* New York: World Book Co., 1946.
19. Lorr, M., & Meister, R. K. The concept of scatter in the light of mental test theory. *Educ. psychol. Meas.*, 1941, **1**, 303–309.
20. Magaret, Ann, & Wright, C. Limitations in the use of intelligence test performance to detect mental disturbance. *J. appl. Psychol.*, 1943, **27**, 387–398.
21. McNemar, Q. *The revision of the Stanford-Binet Scale.* Boston: Houghton Mifflin, 1942.
22. Piotrowski, Z. A. Objective signs of invalidity of Stanford-Binet tests. *Psychiat. Quart.*, 1937, **11**, 623–636.
23. Piotrowski, Z. A. The test behavior of schizophrenic children. *Proc. Amer. Ass. Stud. ment. Def.*, 1937, **42**, 78–90.
24. Rabin, A. I. Test score patterns in schizophrenia and non-psychotic states. *J. Psychol.*, 1941, **12**, 91–100.
25. Rabin, A. I. Differentiating psychometric patterns in schizophrenia and manic-depressive psychosis. *J. abnorm. soc. Psychol.*, 1942, **37**, 270–272.
26. Rapaport, D., *et al. Diagnostic psychological testing*, Vol. I. Chicago: Year Book Publishers, 1945.
27. Roe, Anne, & Shakow, D. Intelligence in mental disorder. *Ann. N. Y. Acad. Sci.*, 1942, **42**, 361–490.
28. Schafer, R. The significance of scatter in research and practice of clinical psychology. *J. Psychol.*, 1944, **18**, 119–124.
29. Schafer, R. *The clinical application of psychological tests.* New York: International Univ. Press, 1948.
30. Schafer, R., & Rapaport, D. The scatter in diagnostic intelligence testing. *Character & Pers.*, 1944, **12**, 275–284.
31. Shipley, W. C., & Burlingame, C. C. A convenient self-administering scale for measuring intellectual impairment in psychotics. *Amer. J. Psychiat.*, 1941, **97**, 1313–1324.
32. Terman, L. M., & Merrill, M. A. *Measuring intelligence.* Boston: Houghton Mifflin, 1937.
33. Thurstone, L. L. *Multiple factor analysis.* Chicago: Univ. Chicago Press, 1947.
34. Wechsler, D. *The measurement of adult intelligence.* Baltimore: Williams & Wilkins, 1944.
35. Wells, F. L. *Mental tests in clinical practice.* Yonkers, N. Y.: World Book Co., 1927.
36. Wittman, P. The Babcock deterioration test in state hospital practice. *J. abnorm. soc. Psychol.*, 1933, **28**, 70–83.
37. Wolfle, D. Factor analysis to 1940. *Psychometr. Monogr.*, 1940, **3**, 1–69.

Received January 12, 1949.

NUMERICAL TRANSFORMATIONS IN THE ANALYSIS OF EXPERIMENTAL DATA[1]

C. G. MUELLER[2]
Psychological Laboratory, Columbia University

This paper is concerned with some problems that arise in the analysis of experimental data and, in particular, with a set of operations frequently useful in such analysis, the numerical transformation. The term transformation, as here used, refers to the operations by which one set of numbers or scores is changed into another set. The procedures of computing logarithms or reciprocals of scores are familiar examples. In what follows we shall consider some of the reasons for converting a set of data into another form and discuss some of the problems that may arise from the application of such a conversion.

Transformations are used most frequently in three situations: (a) the situation in which a quantitative theoretical statement or prediction is reduced to a simpler form so as to permit a more convenient inspection of the agreement between data and theory; (b) the situation in which we seek an empirical equation to describe a set of data; and (c) the situation in which one frequency function is changed to another to allow the application of more efficient statistical tests.

Transformations to Test Theory

The reduction of a quantitative theoretical statement to a simple convenient form for test often begins with an attempt to express the relationship in linear form. [It is important to note that many expressions do not permit such manipulation (see Lipka, **31**).]

1. Consider a theory which demands that, to produce a threshold effect, the product of the intensity, I, of a light stimulus and its duration, t, be a linear function of the duration (**25**); that is

$$It = a + bt, \qquad [1]$$

where a and b are constants. In such a case, a plot of the product, It, against t should give a straight line with a slope of b and an intercept constant a. Since equation [1] may also be expressed in the form

$$I = b + \left(\frac{a}{t}\right), \qquad [2]$$

equation [2] presents an alternative way of testing the agreement between ex-

[1] Prepared under Project No. NR142-404; Contract N6onr-271, Task Order IX, between Columbia University and the Office of Naval Research, U. S. Navy.

[2] The writer wishes to express his appreciation to Prof. C. H. Graham for many critical suggestions as to the form and content of this paper. He is also grateful to Professors H. E. Garrett and Joseph Zubin for their comments on the manuscript.

perimental data and theory: a plot of I against the reciprocal of t should give a straight line with a slope of a and an intercept constant of b. Under these circumstances a reciprocal transformation of t might be used in analysis.

2. As another example, suppose there exists an hypothesis of visual figural after-effects of the following sort.[3] Assume that any symmetrical inspection figure viewed in a given position builds up a region of depolarization, and the extent of the area of depolarization for a given figure varies as a function of the exposure of the inspection figure. With cessation of stimulation by the inspection figure, the area of depolarization decreases in size as a function of time after inspection. The presentation of a test stimulus after removal of the inspection figure involves the interaction of the depolarization due to the test figure with that remaining from the activity of the inspection figure in such a way that the resultant depolarization at a point is the sum of the two depolarizations at that point. The position of an inflection point, L, in the resultant distribution of depolarization is a function of the distribution set up by the inspection figure, and the point of inflection, L, is assumed to vary with time after inspection according to the following equation

$$\frac{d(L - L_0)}{dt} = -k(L - L_0), \qquad [3]$$

where L_0 is the center of the area of depolarization due to the inspection figure, and k is a constant.

We refer our conceptual scheme to observable variables by the assumption that the displacement, D, of a setting, observed experimentally, is directly proportional, by way of a constant, m, to the distance, $(L–L_0)$, i.e.,

$$D = m(L - L_0). \qquad [4]$$

From equations [3] and [4] it follows that

$$\frac{1}{m}\cdot\frac{dD}{dt} = -\frac{kD}{m}; \qquad [5]$$

and on cancelling the m's on both sides and rearranging we have

$$\frac{dD}{D} = -k dt. \qquad [6]$$

On integrating, equation [6] becomes

$$\log_e D = -kt + C \qquad [7]$$

or its equivalent,

$$D = ce^{-kt}$$

where C is the constant of integration, and c its antilogarithm. We may evaluate c by noting that when t is zero, c is equal to D_0, the value of D at the start of the process. Thus,

$$D = D_0e^{-kt}; \qquad [8]$$

[3] The reader should not consider the following account as a serious statement of theory; it is presented to exemplify the use of transformations in theory testing; no claim is made as to its precise value for perception. For theoretical background, see Köhler and Wallach (*Proc. Amer. Philos. Soc.*, 1944, **88**, 269–357.)

and our theory states that the magnitude of the displacement of the equality setting is a negative exponential function of the time interval between the end of the inspection period and moment at which the setting is made. By taking the more convenient common logarithms of [8] we observe that

$$\log_{10} D = \log_{10} D_0 - 0.4343kt. \quad [9]$$

Since equation [9] is linear in form, we may test the agreement between data and theory by plotting $\log_{10} D$ against t and observing whether the experimental points may be represented by a straight line. The slope of such a plot is 0.4343 k and the intercept is $\log_{10} D_0$. Thus a logarithmic transformation of the dependent variable may be used conveniently in the present analysis.

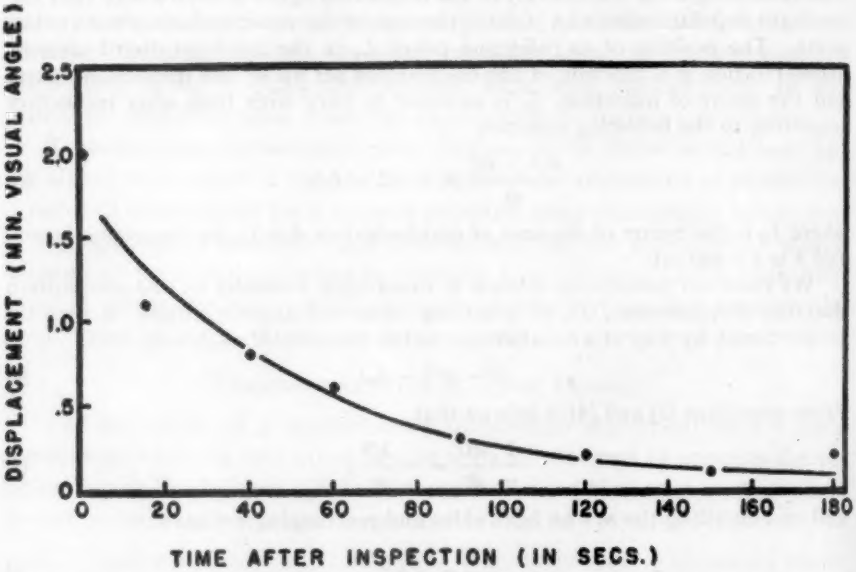

FIG. 1. HAMMER'S DATA [TABLE I IN (24)] ON MEAN DISPLACEMENT OF "EQUALITY" SETTINGS AS A FUNCTION OF TIME AFTER INSPECTION PERIOD.

Hammer (24) has investigated the magnitude of the displacement of setting as a function of time after the inspection period, and Fig. 1 shows a plot of some of her data (her Table I, data for Group Mean, Right Displacements). Fig. 2 shows a semi-logarithmic plot of these same data. With the exception of the point at 180 seconds, the relation in Fig. 2 may be represented by a straight line. The deviant point at 180 seconds is not far removed from the line of fit (calculated from the straight line plot) through the original data of Fig. 1 and lies well within the limit of variability found by Hammer (about 1.4 minutes).

3. Of course, not all problems are as simple as the examples given, and in many cases a major difficulty in testing a quantitative theoretical statement arises from attempts to determine a transformation appropriate to the solution of certain constants. An example of increased complexity in the handling of

transformations is found in a paper by Smith (43) on visual intensity discrimination. The problem was one of testing the agreement between experimental data on human intensity discrimination and Hecht's theoretical account (26).

Hecht's equation relating the adapting intensity, I, and the just discriminable difference in intensity, ΔI, is

$$\frac{\Delta I}{I} = \frac{c}{a^2 k_2}\left[1+\frac{1}{(KI)^{1/2}}\right]^2, \qquad [10]$$

where c, a, k_2, and K are constants. Smith proceeded in the following way. On taking the square root of both sides of [10], the equation may be written

$$\sqrt{\frac{\Delta I}{I}} = \sqrt{\frac{c}{a^2 k_2}} + \sqrt{\frac{c}{a^2 k_2}}\left[\frac{1}{\sqrt{K}}\cdot\frac{1}{\sqrt{I}}\right],$$

or, in its equivalent form,

$$\sqrt{\frac{\Delta I}{I}} = C + \frac{C'}{\sqrt{I}},$$

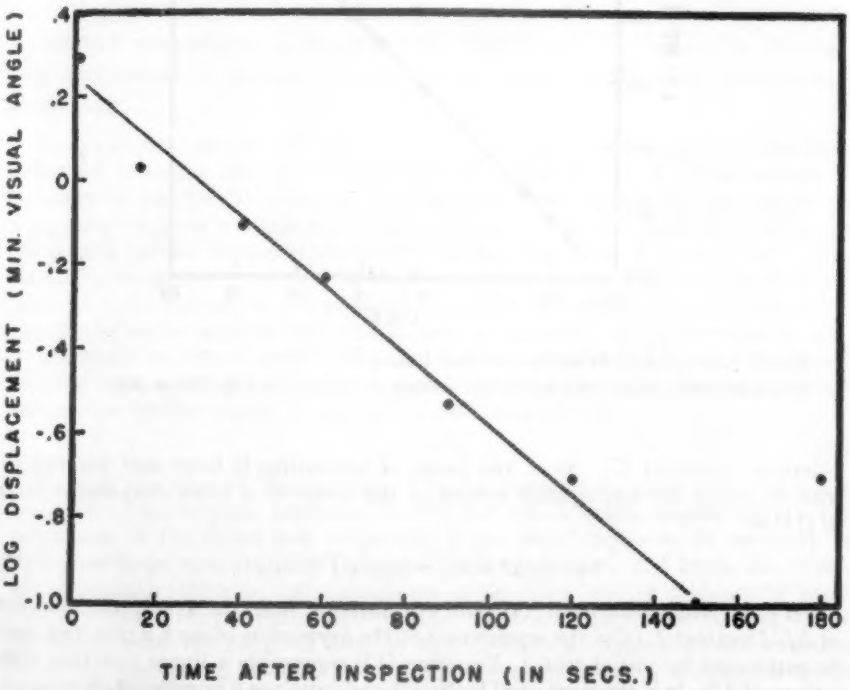

FIG. 2. COMMON LOGARITHM OF THE MEAN DISPLACEMENT AS A FUNCTION OF TIME AFTER INSPECTION PERIOD.

Data from Figure 1.

where

$$C = \sqrt{\frac{c}{a^2k_2}} \quad \text{and} \quad C' = \frac{C}{\sqrt{K}}.$$

Multiplying both sides of the latter equation by the square root of I leads to the expression

$$\sqrt{\Delta I} = C\sqrt{I} + C'. \qquad [11]$$

By this development, it is shown that theory demands that a plot of the square root of ΔI against the square root of I should give a straight line of slope C and

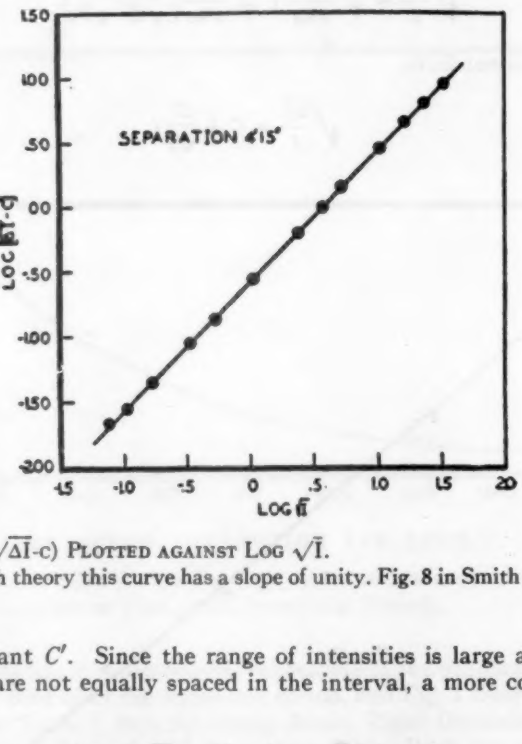

FIG. 3. LOG $\sqrt{\Delta I}$-c) PLOTTED AGAINST LOG $\sqrt{I}$.
In accord with theory this curve has a slope of unity. Fig. 8 in Smith (43).

intercept constant C'. Since the range of intensities is large and the experimental points are not equally spaced in the interval, a more conveinent form of [11] is

$$\log_{10} (\sqrt{\Delta I} - C') = \log_{10} \sqrt{I} + \log_{10} C, \qquad [12]$$

when C' is determined from [11] after evaluating C from the asymptote of a plot of $\Delta I/I$ against I. (C is the square root of the asymptote of such a plot and may be estimated by visual trial.) Equation [12] represents a linear function with a slope of 1.0. It is the form used by Smith and involves a number of elementary transformations, including subtracting a constant and taking the logarithms and square roots of numbers. Fig. 3 shows the agreement between Smith's data and Hecht's theory.

Transformations to Aid Description

In some cases transformations may be applied for the sole purpose of finding an approximate descriptive expression for the relations between variables. A problem may, for example, demand a quantitative statement of the approximate relationship between two variables for purposes of designing equipment, controlling certain forms of behavior, and the like. As opposed to the situation discussed in the preceding section, such a quantitative expression would, in general, have no theoretical basis, but would describe the data within the limits set by the consistency of the data and considerations of simplicity involved in the selection of an expression.

One method of arriving at an empirical formula for any particular set of data is to apply various transformations to one or both of the variables until a set of manipulations is found which yields a straight line. The transformations finally selected as successful in this respect may provide an empirical statement of the way in which the original variables are related. It is important to note, of course, that these computational procedures alone imply no theory, and the resulting descriptive expression in no way constitutes a theory of the data under consideration.

1. An experiment on the relation of area of a visual stimulus to the absolute threshold provided the data represented in Curve A, Fig. **4**, where values of intensity of the visual stimulus (at threshold) are plotted against values of area, selected from a much wider range (Graham, Brown, and Mote, **20**). In our search for an empirical expression to describe these data we may systematically investigate the simpler and more common transformations with respect to their success in giving linearly related variables. In the present instance the curve appears hyperbolic, and a reciprocal transformation of one of the variables (area) results in a straight line relationship with an intercept value of zero, as shown in Curve B, Fig. 4. This finding, of course, implies that the product of the original variables is a constant, that is

$$A \cdot I = k, \qquad [13]$$

where A is the area of the stimulus, I is the intensity at threshold, and k is a constant. This formula expresses Ricco's law which holds, approximately, for small area in the fovea and periphery; it has been shown to be seriously in error over large ranges of stimulus area (Graham, Brown, and Mote, **20**; Wald, **46**). Equation [13] in no way constitutes an "explanation" or a theory of area-intensity relations; the operations we have performed give it no rational basis. Within a restricted range of data it may provide an approximate correction factor for comparing thresholds of differently sized areas or adequately serve some other practical purpose. Graham, Brown, and Mote have shown that, on rational grounds, an entirely different sort of equation fits the data over a much larger range of stimulus area.

2. Another illustration of the procedure under consideration may be taken

from the data of Schoenfeld, Antonitis and Bersh (41) which are presented in Fig. 5, where the median number of responses (bar-pressing responses in a modified Skinner apparatus) in a one hour test period are shown for six successive days. The animals had no history of reinforcement in this situation and were not reinforced during test sessions. The data therefore represent the "operant rate of responding" on successive days. We guess that the data are exponential

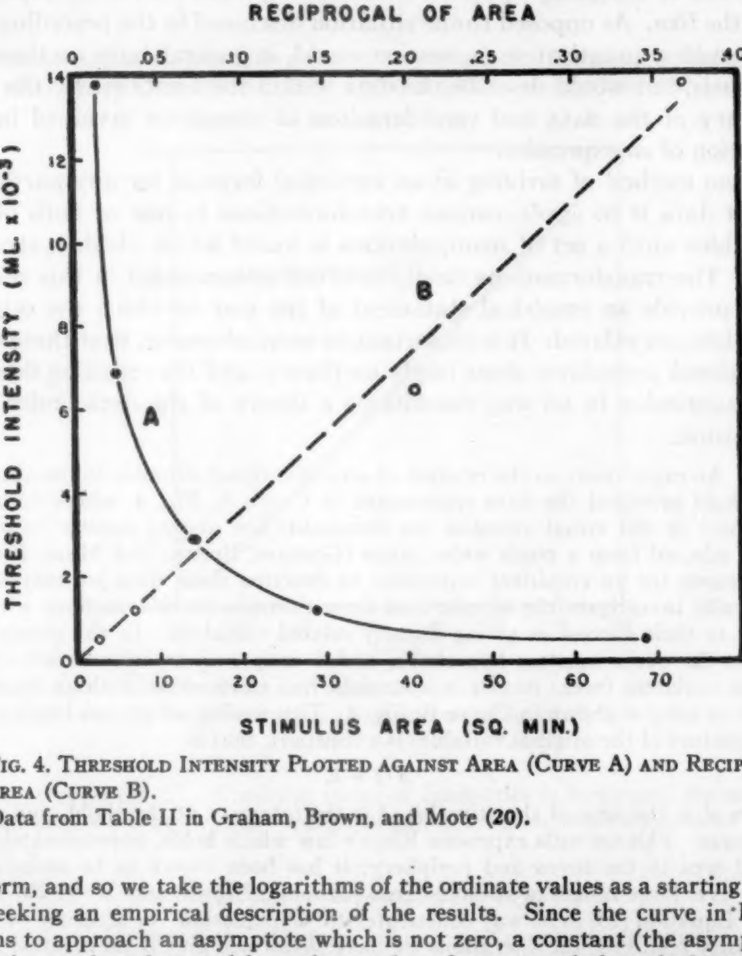

FIG. 4. THRESHOLD INTENSITY PLOTTED AGAINST AREA (CURVE A) AND RECIPROCAL OF AREA (CURVE B).

Data from Table II in Graham, Brown, and Mote (20).

in form, and so we take the logarithms of the ordinate values as a starting point in seeking an empirical description of the results. Since the curve in Fig. 5 seems to approach an asymptote which is not zero, a constant (the asymptotic value) must be subtracted from the number of responses before the logarithms are taken. If we take this asymptotic value to be 9.0, subtract it from each of the median values plotted in Fig. 5 and take the logarithm of the difference, we obtain the results shown in Fig. 6. The data deviate little from linearity which indicates that the original data may be approximately described by an exponential equation. Obviously the equation has significance only for integral

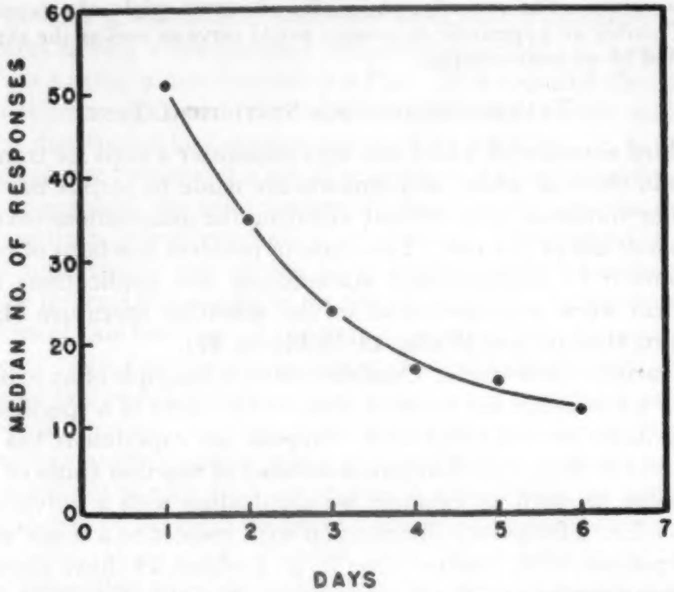

FIG. 5. MEDIAN NUMBER OF RESPONSES IN AN HOUR ON DIFFERENT TEST DAYS. Data from Schoenfeld, Antonitis, and Bersh (41).

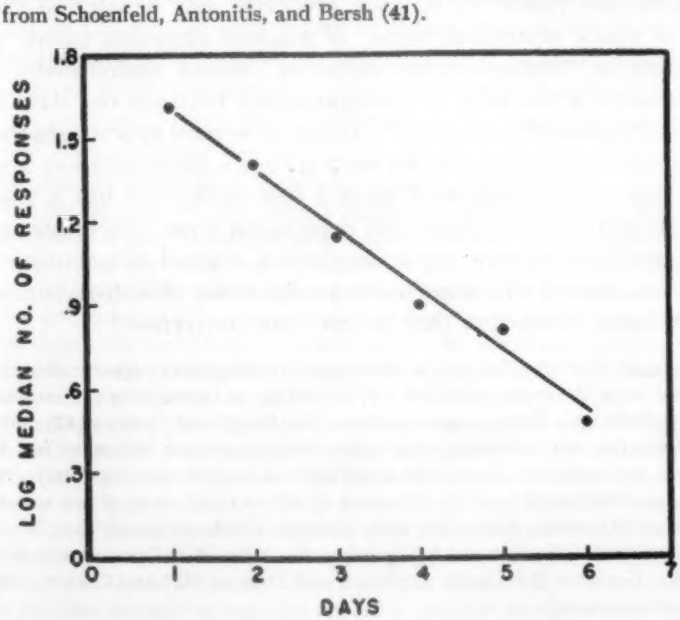

FIG. 6. LOG MEDIAN NUMBER OF RESPONSES IN AN HOUR AS A FUNCTION OF TEST DAY. Data from Fig. 5.

abscissa values. Theoretically, an infinity of curves might also represent the data. Possibly an hyperbolic expression would serve as well as the exponential and would be no more complex.

Transformation for Statistical Test

A third situation in which one may encounter a need for transformations is in the case where adjustments are made to permit the use of a particular statistical test without violating the assumptions involved in the efficient use of the test. This class of problem has been of considerable concern to mathematical statisticians and applications of their theoretical work may be found in the scientific literature (Beall, **4**; Eisenhart, Hastay, and Wallis, **13**; Williams, **47**).

Univariate distributions. Consider a simple example of an experiment in which we measure the reaction time to the onset of a specified visual stimulus under several conditions. Suppose our experiment has yielded several sets of data, distributions of number of reaction times of various magnitudes. In such an example we are dealing with a univariate distribution, i.e., a frequency distribution with respect to a single variable. Since experiments on reaction time (e. g., Jenkins, **29**) have shown that, under many conditions, the frequency distributions of reaction time are not normal but positively skewed, problems may develop in the application of many statistical tests. If marked skewness exists, a direct application of "normal curve statistics" would undoubtedly lead to error because of the lack of correspondence between our data and the conditions imposed by the test.[4] However, several procedures for avoiding this error are open to us: we may (1) use a statistical test applicable to our type of distribution if such a test exists, (2) use a test which makes no assumptions about the population from which our sample is drawn, or (3) transform our scores into a normal distribution so that we may use one of the many tests available for this distribution. It is with the latter procedure that we are here concerned.[5]

[4] The degree of error, of course, is the important thing in any specific situation. If the error is not large the experimentalist will be willing to ignore it in the interest of simplicity of calculation. There is some evidence [see Smith and Duncan (**42**) and Cochran (**7**) for discussion and references] that minor deviations from normality in fairly large samples do not seriously impair the sensitivity of certain statistical tests, but unrestricted generalization of belief in this state of affairs to all cases of non-normality, especially in small samples, is certainly to be avoided. Evidence is clear that, in small samples, serious error is introduced by ignoring the form of the population distribution [Rider (**39**), Festinger (**16**); again see Smith and Duncan (**42**) and Cochran (**8**) for discussion and references].

[5] It is not the purpose of this paper to discuss all of the methods of handling data from populations having non-normal or "unassumable" characteristics, but some ori-

The general problem that confronts us is one of transforming one set of scores having some specified frequency distribution, $y=f(x)$, into another set having a distribution, $y=F(z)$. It is required that we find some function $z=g(x)$ such that when transformations are made from x to z, z is distributed in a manner specified by $y=F(z)$.

Although conversions of a set of scores into any specified frequency function is possible, the most frequently encountered problem is one of making the data approximate the normal curve. The interest in having the data approximate the normal curve arises mainly from the greater amount of information available on the sampling characteristics of the parameters of this function; if the data can be put into the appropriate form, larger resources exist for testing statistical hypotheses.

At least two possibilities are open to us in our search for an appropriate transformation. The first presents us with a solution of the general problem to an approximation which can be made as exact as we please. Several solutions of this type are available and involve an approximation to the desired function by using a power polynomial or series function [cf. Baker (**1**), Fry (**19**), Kendall (**30**)]. These solutions are general and may be used on any continuous distribution. The second possibility available to us is to see whether one of several convenient transformations, such as taking the logarithms or reciprocals of the scores, gives an appropriate and sufficiently precise solution. Logarithms or reciprocals provide an exact solution in those cases in which $z=g(x)$ is a logarithmic or reciprocal function; practically they have been found to

---

entation with respect to the use of transformations may be gained if some of the other methods of handling data are mentioned. Although the emphasis on sampling distributions of parameters of non-normal populations has not approached that given to normal curve functions, work on non-normal distributions [e.g., Festinger (**16**), Paulson (**35**)] may eventually decrease the advantages of recoding scores to test statistical hypotheses by providing analogous tests for non-normal functions. In the present case of a skewed distribution of reaction time scores we might test for the differences between means in two such distributions by using a test suggested by Festinger (**16**). Unfortunately such techniques are not numerous.

The selection of a statistical test which makes no assumptions about the population distribution will probably become an increasingly important procedure to research workers. If we are willing to sacrifice some of the information present in our scores by dealing with ranks or orders of scores, statistical procedures are available for testing many hypotheses, e.g., that population correlation is zero [Hotelling and Pabst (**28**); Pitman (**37**)] and that two samples came from the same population [Wald and Wolfowitz (**45**); Pitman (**36**); Mann and Whitney (**32**); Festinger (**17**); Dixon (**12**)]. In cases where normal curve statistics are appropriate they are more powerful than non-parametric tests, since the latter do not use all the available information; but the generality of statistical tests of the latter type presents a definite advantage to the research worker.

provide sufficiently accurate approximations in a large number of cases encountered experimentally.

If we were interested in approximately "normalizing" our reaction time data, some convenient transformation may serve as a first step. Jenkins (29) has shown that the logarithms of reaction time scores are approximately normally distributed, and this transformation may be our first choice. If we had two distributions on a single dimension (such as two sets of reaction time scores) and were interested in significance of the differences between means, such a transformation would be an appropriate step preliminary to the application of a t-test or some similar test of significance.

Bivariate distributions. In the field of statistical analysis one frequently encounters the variate transformation in situations which involve two variables rather than one.

Transformations, as applied to bivariate distributions, become an important tool of analysis in two cases: (a) in the fitting of theoretical or empirical curves to data, and (b) in the application of tests of significance involving an "analysis of variance." Most applications of "analysis of variance" techniques and most fitting of curves by the method of "least squares" involve the assumptions of normality and of homogeneity of variance in the various columns.[6] If the heterogeneity of variance exists and results from a correlation between mean values and standard deviation values, it is frequently possible to find a simple transformation by which variability may be equalized. Examples of useful transformations in the application of "analysis of variance" tests have been extensively discussed by Bartlett (2, 3), Cochran (6), Curtiss (12), and others. In the present account we consider a few examples taken from the field of psychology.

1. Many groups of data in psychology show a variability which is approximately proportional to the mean. For example, the measurement of differential thresholds in sensory psychology provides data demonstrating a proportionality between the mean and standard deviation over a considerable range of the controlling variable, intensity. If we attempt to analyze the effect of certain variables on mean differential threshold by using "analysis of variance" techniques we require some method for equating the variances in the several classes. In such a situation it may be shown that the variability is *roughly* equalized by a logarithmic transformation.

Consider an example in a visual intensity discrimination experiment of finding a transformation such that the variability of ΔI is constant for various values of ΔI. Crozier (10) has shown that the standard deviation of ΔI is approximately proportional to ΔI, that is,

[6] Exceptions occur in "analysis of variance" tests of the non-parametric type [Pitman (38)] and in "least squares" solutions utilizing a weighting term, as we shall see below.

$$\sigma_{\Delta I} = k\Delta I, \quad [14]$$

where k is a constant. We require a transformation to y, where $y = f(\Delta I)$, such that the standard deviation of y is constant for all mean values of y. The variance of a function of a variable can be represented *approximately*[7] (30) by the equation

$$\sigma_z^2 = \sigma_x^2(dz/dx)^2, \quad [15]$$

where x is the variable and $z = f(x)$. From [14] and [15] it follows that

$$\sigma_y = k\Delta I[dy/d(\Delta I)], \quad 16]$$

where $y = f(\Delta I)$. If we let σ_y equal a constant, a, and transpose, we obtain

$$dy/d(\Delta I) = a/k\Delta I. \quad [17]$$

The function, $y = f(\Delta I)$, which has the latter derivative is $(a/k) \log_e \Delta I$; and so logarithms of ΔI are the appropriate numbers to use in this type of experiment when homogeneity of variance is presupposed.

2. Sensory psychology is not the only area from which examples may be drawn; many studies in conditioning and learning, for example, present similar problems. As an example taken from the field of conditioning we may examine the data of Zeaman (48) on the changes in latency of a running response during acquisition.[8] Zeaman used a straight runway of the Graham-Gagné type and included in his measures the time interval between the opening of the door to the runway and the entry of the animal onto the runway. An examination of the raw data shows (1) that the distribution of latency scores for a group of animals on any particular trial is markedly skewed and (2) that the variability of each distribution varies with the mean value.

Assume that we are interested in arriving at a "best fitting" line to represent the data or in testing the hypothesis that there are no significant changes in average latency from trial to trial. Most methods of satisfying these interests require that the distribution of scores for each trial be normal and that the variability be constant from trial to trial. This is not the case with the latency scores, so we seek some transformation that will yield scores satisfying these conditions. We hope that the transformation that stabilizes the variability in the various trials also normalizes the distributions for each trial. In view of the success of the logarithmic transformation in approximately "normalizing" some reaction time data, we may begin our analysis by taking the logarithms of the latency measures. To test the success of this device in performing the de-

[7] Equation [10] and the development that follows involve several severe approximations; they are presented for their heuristic value to those not familiar with the argument. For some difficulties with this development and a more rigorous account of the transformation discussed here (and several others) see Curtiss (11).

In addition to the approximations involved it is important to note that the operations outlined here provide no guarantee of normality even though they purport to stabilize variability. In those cases where we are dealing with particular forms of skewness, such as Poisson distribution, logarithmico-normal distribution, etc., the transformation which stabilizes variability will also normalize the data. Forms of variance heterogeneity are imaginable, however, in which this would not be the case.

[8] The author wishes to express his gratidue to Dr. David Zeaman for making the data available but assumes full responsibility for any analyses or conclusions that result from their consideration here.

sired function let us cumulate the percentages of cases falling above specified log latency values and plot this cumulated percentage on probability paper with log latency on the abscissa. The cumulative plot of a normal distribution should give a straight line whose slope is proportional to the standard deviation. Such a plot of Zeaman's data is shown in Fig. 7. The data for the odd numbered trials are shown. The approximate normality of the distributions of scores for each trial is obvious from the fact that the points may be approximately represented by straight lines. The homogeneity of variance is demonstrated by the fact that

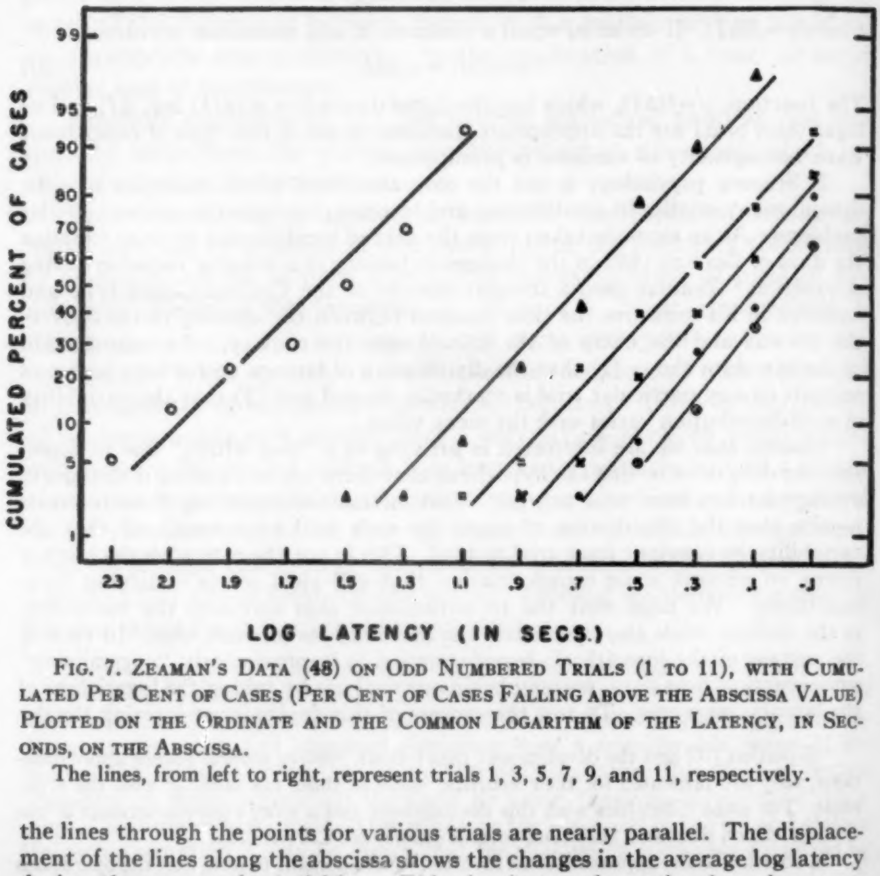

FIG. 7. ZEAMAN'S DATA (48) ON ODD NUMBERED TRIALS (1 TO 11), WITH CUMULATED PER CENT OF CASES (PER CENT OF CASES FALLING ABOVE THE ABSCISSA VALUE) PLOTTED ON THE ORDINATE AND THE COMMON LOGARITHM OF THE LATENCY, IN SECONDS, ON THE ABSCISSA.

The lines, from left to right, represent trials 1, 3, 5, 7, 9, and 11, respectively.

the lines through the points for various trials are nearly parallel. The displacement of the lines along the abscissa shows the changes in the average log latency during the course of acquisition. This simple transformation has given us a function in Y (log latency) and X (trials) which has the characteristics that, for changes in X, we obtain changes in the mean value of Y but no appreciable changes in the variability of Y or in the form of the distribution of Y values. This result confirms a much earlier report by Graham and Gagné (21) on the advantages of using the logarithms of the latency measures in a similar situation.

As has been frequently emphasized in the statistical literature, if a group

of scores is normally distributed along some dimension, then scores along any non-linear function of this dimension will not be normally distributed. In the present instance the fact that the logarithms of the latency scores are distributed normally means that the latency scores themselves are not normally distributed. A plot, on probability paper, of the cumulative frequency of the latency scores against the magnitude of the latency shows a markedly curved line which shows deviation from normality, and the degree of curvature varies with mean value or position on the probability grid. The existence of such information, of course, argues against the application of normal curve statistics to the raw scores.

3. The examples considered so far by no means exhaust the areas in which transformations are, or should be, an important step in analyzing experimental data. Space permits only brief mention of some other examples. An important paper by Morgan (**33**) shows the advantage to be gained by transforming data obtained in experiments on food hoarding in rats. Morgan shows that the logarithms of the number of pellets horded are distributed more normally than are the numbers themselves and that variance under the various conditions is more nearly equal with transformed scores.

Haggard (**22**) and Haggard and Garner (**23**) have considered the problem of the transformation of experimental data in some detail in a discussion of the analysis of the results on the galvanic skin response. Haggard and Garner show that the variability of the GSR measures is approximately proportional to the mean value of the GSR, a consideration which, as we have seen above, leads to a logarithmic transformation if we seek scores showing equal variability. In order to scale the GSR measures, Haggard and Garner employ an additional computation which need not concern us here.[9]

[9] Exactly how the transformation finally selected equalizes variability (in addition to equalizing the response magnitude to a constant stimulus) is not clear. Haggard and Garner present data on (1) the magnitude of the galvanic skin response as a function of the resting resistance and (2) the variability of the GSR. The data on the magnitudes of the response as a function of resting resistance is represented by an equation of the form

$$\Delta R = ae^{bR}, \qquad [a]$$

where ΔR is the change in resistance when the GSR is measured; R, the level of resistance just before the response occurs; a and b, constants; and e, the base of Napierian logarithms. From [a] it follows that

$$\log_e \Delta R = \log_e a + bR \qquad b]$$

and that

$$\frac{\log_e \Delta R - \log_e a}{R} = b. \qquad [c]$$

The left side of equation [c] indicates the transformation used. Haggard and Garner summarize the data on the variability of the response by the statement that the standard deviation of the response magnitudes is approximately proportional to the mean value. We have seen above that under such circumstances the logarithms of the ΔR values should therefore show equal variability. If we then divide these log values by a variable term, R, as dictated by equation [c], variance heterogeneity should result. Computations by Haggard and Garner indicate that this is not the case.

SOME THEORETICAL CONSIDERATIONS

Transformations required by theory. The selection of the appropriate transformation to "normalize" the data or to equalize the variability in the cases discussed so far has been on the basis of empirical considerations; we have been interested in satisfying certain conditions, as, for example, the requirement that the transformed scores be normally distributed and/or show homoscedasticity. On this basis we sought transformations which would serve our purpose. Obviously, there are many transformations which will satisfy any specified condition to any level of approximation we may designate, and our selection from among these possibilities has been on the basis of simplicity, ease of calculation, and degree of approximation desired. In certain instances, however, other considerations may enter to determine the form in which the data may be most fruitfully analyzed, and it is here that we encounter the role of scientific theory in restricting the range of selection of the appropriate dimension for analysis.

Earlier it has been shown that, where there is evidence that the magnitude of variability is roughly proportional to the mean value, we may approximate homogeneity of variance by making a logarithmic transformation of the data. Frequently we are led to make such a transformation on the more indirect basis of a successful theoretical formulation in the field with which our experiment is concerned, even though our data may not contain conclusive evidence that variance heterogeneity would occur without such a transformation.

1. Consider the case in which changes in some dependent variable are a function not only of the changes in the independent variable but also some function of the magnitude of the dependent variable. A model for such a relationship may be found in many photochemical systems, for example, those in which the amount of decomposition is a function not only of the intensity of light but also the amount of photosensitive material available.

A general form of the kind of theoretical statement frequently encountered in photochemistry is

$$-dy/dt = kIf(y) \qquad [18]$$

where y is the amount of photosensitive material, I is the light intensity, k is a constant, and $f(y)$ is usually y or y^2. Rearrangement of [18] yields

$$-kIdt = dy/f(y),$$

from which it follows that

$$-kIt = \int dy/f(y). \qquad [19]$$

It may be shown (Cramer, 9) that, if variations in the left member of [19] can be represented as the activities of independent random variables then, in

the limit, the integral of the right member can be represented in the same way. If we take, as an example, the simple case in which $f(y)$ is equal to y, that is, the rate of change is proportional to y, then the right side of equation [19] reduces to $\int dy/y$ or $\log y$. Thus, if variations in It follow a normal Gaussian function, we should find that the logarithms of y values would be similarly distributed, and analysis involving y must consider this possibility.

2. Although psychology does not provide many examples of transformations which are derivable from theoretical considerations, some cases may exist. Let us reconsider the food hoarding data discussed by Morgan (33). It will be recalled that Morgan found a logarithmic transformation to be useful in treating his data. Assume that, for any experimental condition, i.e., one representing constant drive level, illumination conditions, amount of infant feeding deprivation, etc., there exists a fixed rate of occurrence of the hoarding response and that the responses under any fixed experimental conditions are distributed randomly in time. For events distributed randomly in time, the probability (P_n) of obtaining n events in a time interval, t, is

$$P_n = \frac{(rt)^n e^{-rt}}{n!}, \qquad [20]$$

where r is the rate of occurrence of the events and e is the base of Napierian logarithms. Thus, the number of events observed in intervals of fixed length (experimental test interval) will be distributed according to the Poisson law. One characteristic of the Poisson distribution is that the mean is equal to the variance, which means that any condition which changes the mean thereby changes the variability. By reasoning similar to that used in finding the transformation for the case where the standard deviation is proportional to the mean, we find that the transformation that approximately equalizes the variability of a set of Poisson distributions is the square root transformation. [Again see Curtiss (11) for a rigorous proof.] Thus, from the assumption that the responses are randomly distributed in time for any experimental test condition, we may employ the square roots of the number of hoarding responses as data in situations where homogeneity of variance is presupposed. The author knows of no test of this notion, but presents it as an example of the kind of consideration which may enter.

The reduction of "laws" by the use of transformations. Although the rationale of most deliberate attempts to stabilize the variability of one variable, Y, for various values of another variable, X, may be found in the application of statistical tests, the demands of statistical tests may not be the only, or even the most important, reason for seeking homogeneity of variability. One characteristic of a bivariate frequency function with normal distributions of equal variability in the various columns has not been emphasized in the literature. When the values of one variable, Y, are such that they are normally distributed with equal variability for all values of X, the function relating the two variables, Y and X, does not change in form under translation of the measure of position, i.e., under a change of the representative value employed

(mean, median, mode, centiles, etc.). Although there are statistical bases for choosing among the many available measures of position, these bases may not appear compelling to the theorist if other considerations of importance enter, such as the observation that different functional relations between variables are obtained depending on the measure used. For example, the mean may be rationalized as the most appropriate measure of position *for the normal distribution* because it is the most efficient statistic, i.e., no other statistic has a smaller sampling variability. The differences in efficiency, as for example between the mean and median, however, are hardly of a magnitude to dictate the particular form of analysis if complexities of theory result. When the sets of Y values for various X values yield normal distributions of equal variability no problems of the multiplicity of functional relations arise, since the mean, median, and mode have the same value and the selection of any centile or other fixed position in the distribution merely results in a displacement of the function on the ordinate, or Y, scale by some specifiable amount. The correlation between the mean and standard deviation in cases of skewed distributions, however, guarantees heterogeneity and greatly increases the likelihood of actually finding a multiplicity of functional relations upon changing the positional measure. Thus, if our original data are skewed we could show, for a single experimental condition, as many "empirical behavioral laws" as there are measures of position for a distribution.

It seems highly unlikely that theoretical psychologists will find it fruitful to attempt to formulate the infinite set of theoretical statements necessary to handle all the "laws" that would result from any changes in the measure of position or central tendency. In the case of skewed heteroscedastic distributions it would seem much more feasible to provide a rational account of the transformation required to yield a homogeneous set of normal distributions.

An example of the difficulties encountered when dealing with non-normal distributions is available in the study by Felsinger, Gladstone, Yamaguchi, and Hull (**15**) on reaction latency. Curve A in Fig. 8 shows the cumulated distribution of reaction latencies presented by these authors plotted on probabilty paper. The distribution is obviously skewed (indicated by the marked curvature) and confirms the reports of other investigators (Graham and Gagné, **21**; Zeaman, **48**). In the light of the skewness of the data, it is not surprising that the authors find different functional relations between reaction latency and number of reinforcements if the mean or the median of the distribution of latency scores is taken as the representative value. As in the case of Zeaman's latency data in Fig. 7, we find that the logarithms of the time measures are *approximately* normally distributed as shown by the linearity of Curve B in

Fig. 8, which represents a cumulative percentage plot on probability paper with log latency values on the abscissa.[10]

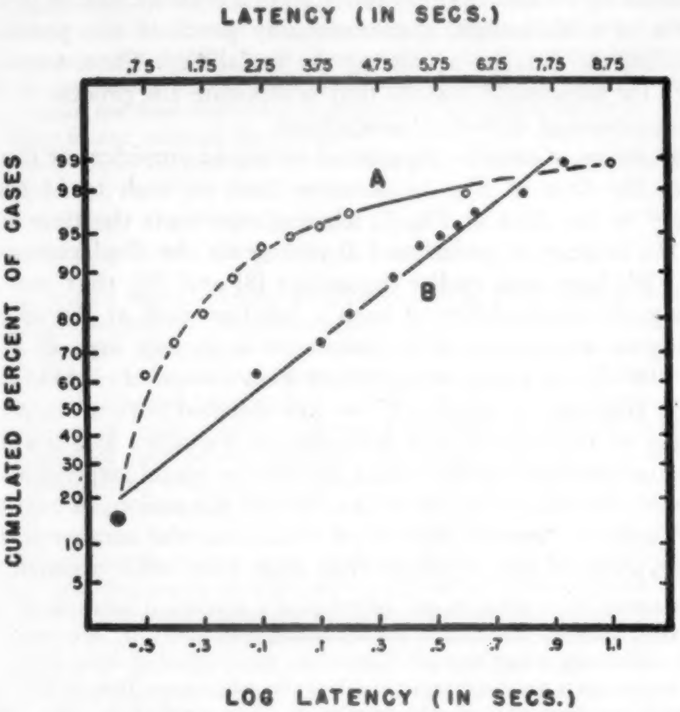

FIG. 8. CUMULATED PER CENT OF CASES (PER CENT OF CASES FALLING BELOW ABSCISSA VALUE) AS A FUNCTION OF THE LATENCY (CURVE A) AND THE LOGARITHM OF THE LATENCY (CURVE B).

Data from Felsinger, Gladstone, Yamaguchi, and Hull (15).

[10] If we are dealing with a logarithmico-normal distribution, the arithmetic mean would not have its usual advantage of statistical efficiency, since in this case the geometric mean shows a smaller sampling variability.

Since we have attempted to solve several problems by the use of logarithms, it is important to emphasize that the present discussion must not, under any circumstances, be interpreted as holding that the logarithmic transformation is a panacea for the experimentalist's analytical problems. The logarithmic transformation does have wide applicability, but obviously the selection of a given transformation in any situation must be made on the basis of merit.

In many of the areas from which we have drawn examples, there is evidence that other simple transformations may satisfy our requirements or that additional conversions will give closer approximations to normality. For example, there is some evidence that log reaction time scores retain some of the positive skewness of the original distributions and that a log-log transformation gives a better approximation to normality. This may also be true where latency measures are employed. In addition, deviations of the sort

Some Consequences of the Use of Transformations

Frequently we find that the selection of a transformation to simplify the form of a theoretical statement may preclude the possibility of normalizing the data or equalizing the variability. Thus, transforming our data for theoretical reasons may complicate the process of meeting the assumptions of statistical procedures.

Curve fitting. Consider the case of fitting an empirical or theoretical curve to the data of Fig. 1. Assume that we wish to fit the curve $D = D_0 e^{-kt}$ to the data in Fig. 1, where t represents the time since the end of the inspection period and D represents the displacement of the setting. We have seen earlier (equations [8] and [9]) that one method of testing the applicability of such a function and, at the same time, providing an evaluation of its constants is to plot $\log_{10} D$ against t. Such a plot should give a straight line with a slope of $-.4343\ k$ and an intercept constant of $\log_{10} D_0$. If we are satisfied with an approximate indication of the relationship between the variables and a subjective estimate of the degree of fit, fitting the line by visual inspection satisfies our needs. For the purposes of the present discussion, however, let us consider a more "exact" method of evaluating the constants involved and note some of the problems that arise from such a consideration.

Our first problem arises in the selection of a statistical solution of the "best fitting" line. Several techniques for obtaining this "best fit" are available, but we shall arbitrarily select one for discussion, the method of least squares. This method serves as a good example and has the advantage that it is widely employed and has been extensively developed. The method is characterized by the assumption that the "best fitting" line is that line about which the weighted squared deviations of data from the line give a minimum sum.

---

shown in the first trial in Fig. 7 may prove to be replicable. For the special cases of the logarithmico-normal distribution where the lower limit of the scores is not zero, the subtraction of a constant (which is the lower limit of the scores) from each score before the logarithms of the numbers are computed may be necessary to insure normality. This consideration arises from an inspection of the parameters of the generalized form of logarithmico-normal frequency function, which is (Cramer, 9),

$$\frac{1}{\sigma(x-a)\sqrt{2\pi}}\, e^{-[\log_e(x-a)\,-\,m]^2/2\sigma^2} \qquad [d]$$

where σ is the standard deviation, and m, the mean of the log distribution; x, a score; a, a constant; and e, the base of Napierian logarithms. The adequacy of the simple logarithmic transformation in any of these cases will depend upon the degree of approximation required as well as the particular characteristics of the data.

In addition to this matter of emphasis, it is to be noted that many useful transformations, for example the arcsin transformation (Zubin, 49; Eisenhart, Hastay, and Wallis, 13), have not even been mentioned because of limitations of space.

The problem of the appropriate weighting of points is one of the important considerations in curve fitting, although it is one which arises only implicitly in most applications of the least squares method. We know that the combination of several distributions of raw scores into a single distribution effectively weights the separate distributions in proportion to their variability; that is, although we ostensibly weight each distribution equally by merely adding the scores to obtain the final distribution, implicit weighting arises from differences in variability in our original distributions. We shall see that this factor plays an important role in the solution of best fitting lines in the cases of original and transformed data, since the usual normal equations

$$aN + b\Sigma x = \Sigma y$$
$$a\Sigma x + b\Sigma x^2 = \Sigma xy \qquad [21]$$

make no provision for differential weighting; that is, they assume homoscedasticity or equal variability along the fitted line. The use of such equations in fitting a line to the logarithms of the scores yields a line around which the sum of the squares of the logarithms of the scores is a minimum. If the variability of Y is equal on a logarithmic scale, i.e., if log Y values show equal variability for all values of X, each point on the curve is equally weighted. But if the Y values themselves show equal variability for various values of X, using equation [21] differentially weights each point as a function of its Y value. Thus the solutions of a and b obtained by fitting a straight line to the transformed data by the least squares formula [21] differs from the solutions obtained by fitting a curve directly to the raw data on the basis of the least squares criterion. In the present case this means that the numerical values of D_0 and k obtained from the data of Fig. 1 depend upon how we fit the exponential curve: directly to the data, or on the basis of a straight-line plot of $\log_{10} D$ against t, using equation [21].

It seems clear that the choice between the constants obtained in the one case by fitting a straight line to the transformed data by equation [21] and in the other case by direct application of least squares to the exponential form depends upon theoretical and empirical considerations; these considerations must be specifiable if the solution of the constants in the function is to have precise significance. If the least squares formula without a "weighting" term is applicable to the untransformed scores, it is not applicable to the transformed scores and vice versa.

It seems imperative that considerations of the sort just outlined be seriously entertained by workers for whom the exact numerical value of constants is important. Variability factors as well as normality factors are too important for the solution of these parameter values to expect that we can efficiently estimate population parameters and give any interpretable statement about confidence intervals without paying specific attention to these and related topics.

The generality of our "least squares" procedure is increased if we employ formulae which incorporate a term for weighting each point. The required formulae are given by the equations

$$a\Sigma w + b\Sigma wx = \Sigma wy$$
$$a\Sigma wx + b\Sigma wx^2 = \Sigma wxy \qquad [22]$$

where w represents the weight and is defined as the reciprocal of the variance. The similarities of formulae [21] and [22] are obvious. In addition to the usual

computations, however, equation [22] requires that we have a method for determining the variability of transformed scores when we are given the variability of the original scores.

We have seen that an approximate solution for the variance of a transformed distribution in terms of the variance of the original distribution is given by equation [15]

$$\sigma_z^2 = \sigma_x^2(dz/dx)^2,$$

where z is the transformed score; x, the original score; and σ^2, the variance of z or x scores. Consider a logarithmic transformation. If the variability is small compared with the mean, then the log of the mean of x is approximately equal to the mean of log x, in which case

$$\sigma_z^2 = (1/\bar{x})^2\sigma_x^2. \quad [23]$$

Equation [23] allows us not only to determine weighting terms but it also allows us to evaluate the error in curve fitting introduced if account is not taken of implicit changes in weighting.

The combination of the weighted least squares formula and the general solution for a set of transformed scores greatly increases the generality of the method of curve fitting and permits other than statistical considerations to dictate when and what transformations shall be made. In particular, convenience or simplicity of the form of our theoretical statements may now play a larger part in determining the analysis of agreement between data and theory.

Correlation. Discussion of the least squares method of curve fitting leads to another example of the bivariate distribution in which similar assumptions are made and similar consequences figure. The example is the product-moment correlation. Although a formula for the correlation coefficient may be developed without assuming a normal bivariate distribution, it must be emphasized that such a proof is not a sufficient condition for the extension to all cases of the large superstructure of correlation theory which is associated with the correlation term in the normal bivariate frequency distribution, particularly that portion of correlation theory having to do with sampling, significance tests, etc. It is true that one may compute a correlation coefficient for any group of paired numbers, but the interpretation to be placed on the result of this computation is not always clear. Statements about the sampling characteristics of the correlation coefficient, inferences relating to population parameters and differences between parameters, are, for the most part, limited to normal bivariate distributions.

Our discussion points to a few factors which may be expected to limit the application of correlation techniques to non-normal populations. Two obvious changes are introduced in the normal bivariate dis-

tribution when non-linear transformations are made in one of the variables. The first of these is a change in the regression line; if the raw scores are linearly related, then the transformation of one of the variables will change this relationship. The magnitude of this change will obviously depend on the transformation made. The second change introduced is also of a type mentioned earlier, that of a change in variability of the distributions of the transformed variable. Both of these factors guarantee that at least one of the forms of bivariate distribution (the one involving transformed scores or the one with the raw scores) is not normal, and different values for the correlation coefficient will be obtained by performing similar computations on the two forms.

1. The importance of the form of the bivariate frequency distribution for product-moment correlation between variables is considered in papers by Ezekiel (**14**) and Stephan (**44**). In discussing the importance of linearity of regression in the product-moment correlation, Ezekiel points out that the logarithmic transformation of one variable frequently converts a curvilinear regression to approximate linearity and that product-moment correlations are increased in such cases. In one example, however, Ezekiel showed that such a transformation not only did not increase, but actually decreased, the correlation. The example was one in which the per cent protein in grain was correlated with the per cent vitreous kernels. As pointed out by Stephan the decrease in correlation is not surprising since the logarithmic transformation was not appropriate in this case and merely served to increase the curvature of the line relating the variables. Stephan showed, however, that by transforming one dimenson to per cent non-vitreous kernels (rather than per cent vitreous) and taking the logarithms of these values, the regression became approximately linear and the variance around the straight line fitted by least squares was considerably reduced. The correlation between the original variables was .73 and the correlation after transformation was −.98. Thus, with the same experimental data and without changing the rank order of the scores, the presumed "amount of variance accounted for" was almost doubled.

2. Another example of a case in which the interpretation of experimental data would have been markedly different if an appropriate transformation had not been employed is an experiment by Mueller and Richmond (**34**) on the comparative reliabilities of several methods of measuring visual acuity. The original data, numbers recorded by the experimenter for each subject, were the denominators of the fraction used to classify each size of figure in commercial tests of visual acuity. The numerator for all such fractions is a constant and typically 20. The fraction has the dimensions of a visual acuity and has considerable empirical and theoretical status as a variable in vision. These considerations and the facts that the distributions of the fractional scores for the several tests were approximately normal (very slightly positively skewed) and the regressions between the tests approximately linear, left little doubt as to the appropriate dimensions for the reliability analysis. When the fractional scores were used in computing the test-retest correlations (product-moment), the results were as given in Table I, column *a*. An interest in the importance

TABLE I

TEST-RETEST RELIABILITIES FOR SIX TESTS OF ACUITY.

Column (*a*) shows the results when the acuity scores are used; column (*b*) shown the results when the reciprocals of the acuity scores are used.

Test	*Reliability*	
	a (acuity scores)	*b (reciprocal of acuity scores)*
A	.81	.81
B	.75	.79
C	.69	.86
D	.68	.76
E	.61	.81
F	.56	.65

of the form of the bivariate distribution, however, prompted an analysis of the test-retest correlations using the originally recorded scores, rather than the fraction scores. The correlations in the case of the original scores are presented in Table I, column *b*. Tests for the significance of the differences (using Fisher's *z* transformation) in Table I, column *a*, leads to the conclusion that test A has a higher reliability than tests C and D, at between the 1 per cent and 5 per cent levels, and higher than tests E and F at less than the 1 per cent level. The same significance tests applied to the data of Table I, column *b*, lead to the conclusion that test C has a higher reliability than D at between the 1 per cent and 5 per cent levels and higher than F at better than the 1 per cent level. Obviously, such a result constitutes a dilemma only if we ignore the importance of the frequency characteristics of the data to be analyzed. Actually, we have reasons to doubt the applicability of our test of significance in the case of the untransformed scores. The point to be emphasized is that test-retest correlations on the recorded data would have led to conclusions incompatible with the results of a more appropriate analysis.

SUMMARY

This paper considers several circumstances which may call for the numerical transformation of data: (1) testing theory, (2) finding empirical descriptive expressions, and (3) satisfying conditions of applicability of certain statistical techniques. The application of transformations in each of these circumstances may be straightforward; nevertheless, account must always be taken of the degree of approximation desired or permissible. In addition, situations involving more than the single case of (1), (2) or (3) will frequently call for incompatible transformations. Such a case might arise, for example, in testing a theoretical prediction of linearity of relationship between two transformed variables; here it might be found that non-normality and heterogeneity

of variability around the theoretical line are introduced by the transforming operations. Such a state of affairs may make certain statistical tests inapplicable. Approximate methods are mentioned by which some problems of this sort may be solved. When sufficient information is available, these methods may be of value in determining the magnitudes of desired constants. Little is gained by employing these statistical methods of curve fitting if adequate information is not present.

For the theoretical psychologist many important considerations, other than those concerned with statistical tests, attach to the various kinds of distributions of two variables. Certain forms in which data may be expressed may lead us to as many "laws" as there are measures of "central value" of groups of data. In such cases appropriate transformations serve to reduce the multiplicity of forms of functional relation between variables, and these transformations provide both a subject matter and a tool for the theorist.

BIBLIOGRAPHY

1. Baker, G. A. Transformation of non-normal frequency-distributions into normal distributions. *Ann. math. Statist.*, 1934, **5**, 113–123.
2. Bartlett, M. S. The use of transformations. *Biometrics*, 1947, **3**, 39–52.
3. Bartlett, M. S. Square root transformations in the analysis of variance. *J. roy. statist. Soc., Suppl.*, 1936, **3**, 68.
4. Beall, G. The transformation of data from entomological field experiments so that analysis of variance becomes applicable. *Biometrika*, 1942, **32**, 243–262.
5. Chapman, R. A. Applicability of the *z*-test to a Poisson distribution. *Biometrika*, 1938, **30**, 188–190.
6. Cochran, W. G. The analysis of variance when the experimental errors follow the Poisson and binomial laws. *Ann. math. Statist.*, 1940, **11**, 335–347.
7. Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 1947, **3**, 22–38.
8. Craig, A. T. On the distribution of certain statistics. *Amer. J. Math.*, 1932, **54**, 353–366.
9. Cramer, H. *Mathematical methods of statistics.* Princeton, N. J.: Princeton Univ. Press, 1946.
10. Crozier, W. J., & Holway, A. H. On the law for minimal discrimination of intensities. III. *Proc. nat. Acad. Sci.*, 1938, **24**, 130–135.
11. Curtiss, J. H. On transformations used in the analysis of variance. *Ann. math. Statist.*, 1943, **14**, 107–122.
12. Dixon, W. J. A criterion for testing the hypothesis that two samples are from the same population. *Ann. math. Statist.*, 1940, **11**, 199–204.
13. Eisenhart, C., Hastay, M. W., & Wallis, W. A. *Techniques of statistical analysis.* New York: McGraw-Hill, 1947.
14. Ezekiel, M. *Methods of correlation analysis.* New York: John Wiley, 1930.
15. Felsinger, J. M., Gladstone, A. I., Yamaguchi, H. G., & Hull, C. L. Reaction latency ($_{s}t_{r}$) as a function of the number of reinforcements

(*N*). *J. exp. Psychol.*, 1947, **37**, 214–228.

16. FESTINGER, L. A statistical test for means of samples from skew populations. *Psychometrika*, 1943, **8**, 205–210.
17. FESTINGER, L. The significance of difference between means without reference to the frequency distribution function. *Psychometrika*, 1946, **11**, 97–105.
18. FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. statist. Ass.*, 1937, **32**, 675–701.
19. FRY, T. C. *Probability and its engineering uses.* New York: Van Nostrand, 1928.
20. GRAHAM, C. H., BROWN, R. H., & MOTE, F. A. The relation of size of stimulus and intensity in the human eye: I. Intensity thresholds for white light. *J. exp. Psychol.*, 1939, **24**, 555–573.
21. GRAHAM, C. H., & GAGNÉ, R. The acquisition, extinction, and spontaneous recovery of a conditioned operant response. *J. exp. Psychol.*, 1940, **26**, 251–280.
22. HAGGARD, E. A. Experimental studies in affective processes: II. On the quantification and evaluation of "measured" changes in skin resistance. *J. exp. Psychol.*, 1945, **35**, 46–56.
23. HAGGARD, E. A., & GARNER, W. R. An empirical test of a derived measure of changes in skin resistance. *J. exp. Psychol.*, 1946, **36**, 59–70.
24. HAMMER, ELAINE R. Temporal factors in figural after-effects. *Amer. J. Psychol.* (In press.)
25. HECHT, S. The nature of the photoreceptor process. In C. Murchison (Ed.), *Handbook of general experimental psychology.* Worcester: Clark Univ. Press, 1934.
26. HECHT, S. A theory of visual intensity discrimination. *J. gen. Physiol.*, 1935, **18**, 767–789.
27. HOLZINGER, K. J., & CHURCH, A. E. R. On the means of samples from a U-shaped population. *Biometrika*, 1928, **20**, 361–388.
28. HOTELLING, H., & PABST, M. R. Rank correlation and tests of significance involving no assumption of normality. *Ann. math. Statist.*, 1936, **7**, 29–43.
29. JENKINS, T. N. Facilitation and inhibition. *Arch. Psychol., N. Y.*, 1926, No. 86.
30. KENDALL, M. G. *The advanced theory of statistics.* Vol. II. London: Charles Griffin, 1946.
31. LIPKA, J. *Graphical and mechanical computation.* New York: John Wiley, 1918.
32. MANN, H. B., & WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. math. Statist.*, 1947, **18**, 50–60.
33. MORGAN, C. T. The statistical treatment of hoarding data. *J. comp. Psychol.*, 1945, **38**, 247–256.
34. MUELLER, C. G., & RICHMOND, P. N. (Unpublished.)
35. PAULSON, E. On certain likelihood ratio tests associated with the exponential distribution. *Ann. math. Statist.*, 1941, **12**, 301–306.
36. PITMAN, E. J. G. Significance tests which may be applied to samples from any population. *J. roy. statist. Soc., Suppl.*, 1937, **4**, 119–130.
37. PITMAN, E. J. G. II. The correlation coefficient test. *J. roy. statist. Soc. Suppl.*, 1937, **4**, 225–232.
38. PITMAN, E. J. G. III. The analysis of variance test. *Biometrika*, 1938, **29**, 322–335.
39. RIDER, P. R. On small samples from certain non-normal universes. *Ann. math. Statist.*, 1931, **2**, 48–65.
40. RIDER, P. R. On the distribution of the ratio of mean to standard devia-

tion in small samples from non-normal universes. *Biometrika*, 1929, **21**, 124–143.

41. Schoenfeld, W. N., Antoinitis, J. J., & Bersh, P. J. Unconditioned response rate of the white rat in a bar-pressing apparatus. *J. comp. physiol. Psychol.* (In press.)
42. Smith, J. G., & Duncan, A. J. *Sampling statistics and applications.* New York: McGraw-Hill, 1945.
43. Smith, J. R. Spatial and binocular effects in human intensity discrimination. *J. gen. Psychol.*, 1936, **14**, 318–345.
44. Stephan, F. E. Alternative statements of percentage data in the fitting of logarithmic curves. *J. Amer. statist. Ass.*, 1931, **26**, 58–61.
45. Wald, A., & Wolfowitz, J. On a test whether two samples are from the same population. *Ann. math. Statist.*, 1940, **11**, 147–162.
46. Wald, G. Area and visual threshold. *J. Gen. Physiol.*, 1938, **21**, 269–287.
47. Williams, C. B. The use of logarithms in the interpretation of certain entomological problems. *Ann. appl. Biol.*, 1937, **24**, 404–414.
48. Zeaman, D. Response latency as a function of the amount of reinforcement. *J. exp. Psychol.* (In press.)
49. Zubin, J. Note on a transformation function for proportion and percentages. *J. appl. Psychol.*, 1935, **19**, 213–220.

Received February 23, 1949.

HISTORICAL BEGINNINGS OF CHILD PSYCHOLOGY

WAYNE DENNIS
University of Pittsburgh

Scientific child psychology is commonly said to have begun with the publication of Preyer's *Die Seele des Kindes* (36) in 1882[1] and Hall's *The Contents of Children's Minds* (20) in 1883. Both of these were important contributions. Preyer's general account of mental development in the first four years was based upon careful and detailed observations of his son, to which were added data contributed by others and some comparative facts concerning behavioral development in animals. Hall's publication represented a very extensive attempt to determine the child's familiarity with a large number of relatively ordinary objects and concepts upon entrance into the first grade. This study of children's concepts was oriented, on the one hand, toward a description of the child's thought tendencies, and, on the other hand, toward the educational implications of the limited experience and understanding of the child upon his entrance into school.

Without doubt these two works are important landmarks in the development of child psychology. However, the history of science warns against attributing absolute originality or uniqueness to any contribution. Every development must be the result of a developmental process. The trends which led Preyer and Hall to observe and record child behavior attracted other men as well, and Preyer and Hall were not the first to carry scientific enterprise into the realm of human infancy.

Both of these men were well aware of the contributions of their predecessors. Preyer's book contains many references to the records of earlier observers, although, to be sure, none had left records as complete as was his own. Hall's study was undertaken as a repetition of an earlier study conducted in Berlin by Bartholomäi (3). Hall, like Preyer, did a much better job than his predecessors. The works of Preyer and Hall were outstanding not because they were novel in conception, but because they were executed with great ability and thoroughness.

It is the aim of this paper to survey the observational studies of

[1] The date of publication of this book is sometimes given as 1881. This seems to be due to the fact that Preyer, in prefaces to editions beyond the first, gave 1881 as the date of the first edition, and later writers have accepted the author's statement. Preyer's *preface* to the first edition is dated October 6, 1881, but a copy of the first edition in the Library of the Surgeon General's Office, Washington, D. C., bears a publication date of 1882 and bibliographical sources of that period give 1882 as the date of the book. It seems likely, therefore, that Preyer, in writing later prefaces, erroneously referred to the date of the first preface as the date of publication of the first edition.

normal child psychology up to the time of the publication of *Die Seele des Kindes* and *The Contents of Children's Minds*. It is believed that a survey of this sort will indicate the proper prominence of the two studies which are usually referred to as the first major studies in child psychology but will, at the same time, show that the early development of child psychology was continuous and gradual.

Before beginning this survey, it will be advisable to attempt to define its scope. It undertakes to include only observational records of normal child behavior; it excludes theoretical treatments of the child. Since nearly every theory of human nature involves some consideration of childhood, to include general and theoretical material on the nature of the child would materially change the task which we have set for ourselves. A justification for excluding this material is found in the fact that theories of child nature have been given adequate historical treatment in several histories of philosophy and in histories of education, while no adequate survey of the early observational studies of children is available.

We have excluded also all publications relating to the behavior of abnormal children. The early medical literature on many types of abnormal children such as hydrocephalics, microcephalics, anencephalics, idiots, and cretins is very extensive. Many of the reports contain some general observations about behavior. The orientation of these reports, however, is on the whole medical rather than psychological, and it seems best to limit the present report to psychological studies of normal children.

In organizing an historical survey, a purely chronological arrangement of data may be followed, or a primary division into topics may be made, the historical sequence being followed within each division. In the present instance, a purely chronological treatment would result in the juxtaposition of entirely unrelated studies. It has, therefore, seemed best to subdivide the studies by subject matter before beginning the historical treatment.

The arrangement of the topics is more or less arbitrary. Since biographical records are among the earliest, and are also the most numerous it seems appropriate to place the infant biographies at the beginning of the survey.

Biographical Studies

Tiedemann's account (47) of the behavioral development of his son was the first infant biography to be published. It appeared in 1787. The years covered by Tiedemann's scanty observations were 1781–1784.

Three years of occasional observations were reported within the space of 25 printed pages.

While Tiedemann was the first to publish a record of the development of an individual child, at least two other persons kept such a record at an earlier date than did Tiedemann. The two earlier diarists were Heroard and Pestalozzi. Heroard was a physician in the French court who was assigned to care for the health of the young prince who became Louis XIII. The physician began his duties at the birth of the prince in 1601 and continued them for many years. Heroard kept a journal in which he recorded the chief events in the court, his own activities, and also, in some detail the health and development of his young charge. Thus, while the developmental data concerning young Louis XIII are mixed with other matters, they are relatively complete. Heroard's journal, edited by Soulie and Barthelemy (**21**) was not published until 1868. It has furnished the basis of a recent book by Crump (**7**), entitled *Nursery Life 300 Years Ago.* Although Miss Crump cites some of the behavioral data, her chief interest lies in picturing the life of the court, rather than the development of the dauphin.

Pestalozzi, in 1774, kept for a period a diary in which he recorded his attempts at the education of his four-year-old son and made some record of his son's behavior. None of this diary seems to have been published in Pestalozzi's lifetime (1746–1827). Parts of it were contained in Niederer's *Notes on Pestalozzi* (**32**) whose date of publication was 1828.

Several studies of individual children made in the first half of the nineteenth century lay unpublished for many years after they were completed. Bronson Alcott kept for a while a record of the development of a daughter born in 1831. This record, still unpublished, is probably contained in notebooks left by Alcott (**30**). A brief summary of this record was presented by Talbot (**46**) in 1882.

Charles Darwin's diary of the development of his infant son, begun in 1840, was not printed until 1877 (**10**). Strumpell's longitudinal study of his child (**43**) was kept during the years 1846 and 1847, but did not appear until 1880.

The earliest baby biography to be published in English was by an American, Mrs. Emma Willard. It appeared as an appendix to an American edition of Madame Necker de Saussure's *Progressive Education.* Its date of publication is 1835 (**50**).

Sigismund's book (**41**), which contains some biographical material, appeared in 1856. Four other individual studies antedated Preyer's *Die Seele des Kindes* by one or two years. These reports were by Taine (**45**), Wyma (**51**), Sully (**44**), and Champneys (**5**).

The data just summarized indicate that at least twelve biographical records were kept before Preyer undertook his observations of his son. They indicate also an increase in frequency of records of this type in the years just preceding Preyer's study.

Language Studies

In a sense, the reports to be listed under this heading might also be called biographical investigations, since all of the early researches on infant speech are longitudinal records of one or two subjects. Many of the baby biographies previously cited contain some account of language development. However, none of those mentioned in the preceding section was solely concerned with speech, whereas that topic is almost the exclusive interest of those to be discussed here. Schleicher (**39**), who wrote in 1861, seems to have been the first to record the speech development of a specific child. His report was followed fairly closely by those of Holden (**24**), Pollock (**33**), Vierordt (**49**), Egger (**12**), Schultze (**40**), and Humphreys (**25**). Since each of these was primarily a brief presentation of individual developmental data, further description seems unnecessary here.

Hun (**26**) described in some detail the invention of new words by a 4½-year-old girl who understood English quite well. This new language she taught to her 18-month-old brother.

Normative Studies

Several early investigations may be grouped together because they share the common characteristic of being interested in the typical performance of some age level, or of being interested in determining the average age of children at the onset of some developmental item. The earliest of these normative studies was that of Feldman (**13**), published in 1833. Feldman was a candidate for the M. D. degree. At that time, it was customary to require each candidate to prepare a short dissertation in Latin. Feldman chose to write concerning the normal functioning of the human body, and included under the heading of the normal state of the body in infants, data on the onset of walking and talking. He presumably obtained his data from mothers' reports, although his statement is not clear on this point. He states that he "observed" these events in the case of 35 infants, but it seems scarcely possible that that he could have had the facilities or the time to observe personally the beginning of speech and of upright locomotion in so many subjects. Feldman[2] writes:

[2] I am indebted to my wife for the translation of this work. The original monograph is available in the Library of the Surgeon General's Office, Washington, D. C.

If you inquire at what time children attain the faculty of walking, that thing I have observed in the case of 35 infants. Of these, 6 were able to walk at the 11th and 12th months; however, the rest required 13 months, and some even 18 months or a small part of the 18th month in order that this feat might be accomplished without aid. All concerning whom I have spoken enjoyed the best of health. . . . I have found among the 35 infants that one of them attained the faculty of speaking the first word in the 14th month, 8 in the 15th month, 19 in the 16th month, 3 in the 17th month, one in the 18th month, and one in the 19th month. Therefore, the normal time for first speech in infants would seem to be the sixteenth month.

It will be seen that Feldman's presentation of his data in regard to walking is very incomplete, but it is clear that the majority of his subjects were one year of age or older before walking alone. His distribution of cases for the onset of speaking is more detailed, but he presents data on only 33 of his 35 cases. Nevertheless his study is notable for preceding by many years the next attempt to set up empirical standards for the onset of locomotion and speech.

The first person to apply a uniform behavioral test throughout a large part of the life span was Quetelet (37) whose work appeared in 1835. Quetelet is known for his contributions to methodology in the field of census records, social statistics and vital statistics. However, his interest extended beyond what is usually thought of as comprising these fields, and he presents in his *Physique Sociale* observations on strength of upright pull and strength of grip of the right and the left hands from ages 6 to 60. His data are presented separately for the two sexes. Yearly age groups are employed from 6 to 21 years; there is a 25-year-old group; thereafter his subjects are grouped by decades. In each group, his averages are based upon ten persons of each sex. His data show that males are superior in strength to females at every age level. The data also show the superior strength of the right hand as compared to the left in dextral individuals.

In addition, Quetelet presented data on the incidence of suicide and of crimes at various epochs of life, including the period below ten years. The data are presented by 5-year step-intervals from 10 to 40 years, and by decades thereafter. Especially noteworthy is the high suicide rate of Berlin youth between 15 and 20 years of age. The data for Berlin refer to the years 1818–1824.

The first normative study of the newborn was that of Kussmaul (28), which appeared in 1859. Kussmaul, like Feldman, was a physician. His experiments with various modes of stimulating the newborn led him to believe that all of the sense organs were capable of some degree of function at birth. He gave fairly adequate description of the repertory of responses of the neonate. This pioneer study was followed

by two similar investigations by Genzmer (**19**) and Kroner (**27**), which added but little to the findings of Kussmaul.

Other studies of neonatal behavior may be called normative only by extending somewhat the meaning of norms, yet it is true that these studies were designed to investigate the characteristics of an age period. Hertz (**22**) and Biedert (**4**) were interested in measuring the force of suction, or negative pressure, of the neonate's nursing response and devised apparatus which they used for that purpose. Donders (**11**), Raehlmann and Witkowski (**38**) and Cuignet (**8**) made observations upon the ocular responses of the neonate. Moldenhauer (**31**) tested the hearing of the young infant.

The Contents of Children's Minds

The studies to be reviewed here served as models for Hall's work by the same title, and Hall in his introduction to his investigation gives an excellent account of his predecessors' work.

The first study, which was conducted in Berlin, was reported by Bartholomäi (**3**). While the Berlin investigators described their problem roughly as that of determining the contents of children's minds, the tests or questions which they submitted to their subjects called for several types of performance. One procedure consisted in asking each subject which of a number of common objects, such as a pond, a lake, a running hare, a squirrel in a tree, a hen with chickens, etc., the child had ever seen. In this portion of the study it was undetermined whether the child who had not had these experiences might nevertheless have some comprehension of the phenomena gained from pictures or from verbal descriptions; the child was interrogated only with regard to his direct experience with the objects. On the other hand, some questions which were employed were essentially tests of language comprehension. It was determined, for instance, whether or not the child knew what was meant by a dwelling, a sphere, a cube, clouds, etc. Other questions asked the child to give the name and business of his father. The examining was done by the regular classroom teachers to whom a circular or directions had been sent. Clearly the research procedures were far from ideal.

In this investigation, which was conducted in 1869 and whose results were published in 1870, 2238 children just entering school were questioned in regard to 75 topics. The chief discovery was the remarkable extent to which first graders are ignorant of simple phenomena. Less than 50% knew what was meant by a sunset, only 30% said they had seen a sunrise, only 11% had seen a river, etc. It was found that girls excelled in regard to knowledge of some concepts, boys in regard to others. Children who had attended kindergarten, on the whole, did

better than those who had not attended, a finding which foreshadows an interest of the present day.

Lange (**29**), in 1879, presented similar questions to 500 children entering the city schools of Plauen, Germany, and to 300 children in near-by rural districts. In regard to the concepts which he attempted to test, many of which were of the nature study sort, the country children proved to have a superior knowledge.

G. Stanley Hall's study of Boston children (**20**) was begun in September, 1880. It was first published in the *Princeton Review* for 1883, and was later republished several times. Hall drew up a list of 134 topics about which his subjects were to be interrogated. It is not within the province of this article to describe Hall's study in full. We may note in passing that Hall's contribution in this field was chiefly that of improvement in method. Instead of utilizing ordinary classroom teachers whose methods of examination could not be controlled, Hall employed four kindergarten teachers who were trained to uniform procedures of questioning, and who frequently met with their director for discussion. In Boston, as in Berlin, the percentage of ignorance concerning simple concepts was surprisingly high.

GALTON'S "INQUIRIES INTO THE HUMAN FACULTY AND ITS DEVELOPMENT"

That Galton's work is not usually given a prominent place in the development of child psychology is probably due to the fact that this field was not his primary interest. It is true, nevertheless, that he should be credited with several contributions to this area of specialization. Although several of his writings bearing upon childhood were included in the *Inquiries Into the Human Faculty and Its Development* which was published in 1882, their first date of publication in most instances was earlier than that year, since Galton's book consisted for the most part of reprints of articles that had previously been presented in scientific and popular journals.

In 1875 Galton (**14**) published the first scientific study of the psychological development of twins, thus initiating one of the major types of investigations of nature and nurture. Galton's method consisted in inquiring concerning twins who in childhood were very much alike to determine whether or not they became more unlike after they left their homes and entered different environments. Conversely, he sought information concerning twins who in the earliest years were exceedingly different, to learn whether or not under the continued influence of the same nurture they became more similar. His materials were obtained

by sending "circulars of inquiry" to persons who were either twins themselves or were relatives of twins.

Galton found in regard to these twins who were remarkably alike in childhood that in some cases "the resemblance of body and mind continued unaltered up to old age, notwithstanding very different conditions of life." Other cases showed divergence with age. He felt that the continued close similarity of some cases despite different environments showed that the divergences of other twins were due to illnesses or to different original endowments which only became apparent in maturity. The second group of data, derived from twins who were originally dissimilar, likewise led Galton to stress the importance of nature, since he found that no matter how similar the environment of the two, they remained markedly different. Galton wrote: "There is no escape from the conclusion that nature prevails enormously over nurture when the differences of nurture do not exceed what is commonly to be found among persons of the same rank of society and in the same country."

In Galton's "Psychometric experiments," first presented in 1879, (**15**), he employed himself as subject in studying the association of ideas. It may seem odd, therefore, to cite this research as a contribution to child psychology. One of Galton's tasks, however, was to attempt to determine when each association had been formed. In many cases this was not difficult, since the date of formation of the association could be established with certainty. Galton found that of 124 associations, 39% of them were formed in boyhood and youth, 46% in adult years, and only 15% could be attributed to recent events. On the basis of these results Galton called attention to the importance of childhood in the formation of associations between ideas.

"Statistics of mental imagery" was the title of a report by Galton in 1880 (**16**). In pursuing his interest in imagery, Galton, as is well-known, devised rating scales for several aspects of imagery. Visual imagery was rated by 100 English men of science and by 172 Charterhouse school boys. Vivid imagery was found to be much more common among the boys, and boys were more able also to call up colored images. Galton believed that the habit of abstract thought resulted in the suppression of visual images.

The study of imagery was pursued further by Galton in an investigation of number forms (**17**). He found that not uncommonly a person located each number within a certain region of his visual space. Most of his informants were certain that their particular forms of visualizing numbers had been present, unchanged, from early childhood. This finding led Galton to ask teachers to question their pupils concerning

this matter. Among subjects of about 14 and 15 years of age, Galton found that clear-cut number forms appeared in about one case in 20 in boys and more often in girls.

On the basis of the number of contributions to the study of mental development which are contained in *Inquiries Into the Human Faculty and Its Development*, it is our belief that this publication should be grouped with those of Preyer and Hall as one of the important early publications in child psychology.

Miscellaneous Studies

In addition to the studies previously discussed, which were capable of grouping, a few remain, each of which stands alone within the epoch under consideration. We shall mention these briefly, following the chronological order.

Miss Elizabeth Peabody (1) in 1835, 1836 and 1837 published accounts of the procedures of Bronson Alcott in his remarkable Boston school. Alcott stimulated student particiption in the educational situation by holding "conversations" with children. Miss Peabody recorded, as nearly verbatim as possible, many of these conversations. The result was a detailed account of child reasoning and child response in Alcott's school.

Darwin's "Expression of the Emotions in Animals and Man" (9) contained some observational material on emotional expression in children. Many of these observations were made on his own children; we have seen that his interest in child behavior began as early as 1840. Some observations were gathered from friends and correspondents.

Clarke (6) described certain images which commonly appear in childhood but which in most cases disappear with age. It seems likely that he was referring to what are now known as eidetic images.

Sikorsky (42) was the first to study decrements of performance during the school day. Investigations of this type later came to be known as studies of mental fatigue.

Vierordt (48) devised a shoe which would record the contact of the feet with the walking surface, and pointed out differences between infantile and adult walking patterns.

Hicks (23) devised a method of recording prenatal movements by measuring displacements of the abdominal wall.

Professional Affiliations of Early Research Workers in this Field

It may be of interest to note the professional affiliations of those who first contributed to our knowledge of child behavior. The occupations of some of the persons to whom we have referred could not be ascertained from the sources available to us. However, it is obvious that physicians were the most frequent contributors to the as-yet-unselfconscious science. Among the members of the medical profession mentioned

in our bibliography are Biedert, Clark, Cuignet, Feldman, Genzmer, Héroard, Hicks, Hun, Kroner, Kussmaul, Moldenhauer, and Wyma. Educators are next in frequency. This professional group includes Alcott and Peabody, Bartholomäi, Lange, Pestalozzi and Strumpell. Vierordt was a physiologist; Darwin a biologist; Quetelet was many things, but was primarily a statistician; Taine and Sully both were philosopher-psychologists. Galton contributed to so many fields that he can scarcely be claimed by any. Preyer was a physiologist. Hall belongs to both psychology and education.

Discussion and Summary

We have reviewed forty-two publications which appeared before 1882. It will be noted that the earliest publication referred to in the preceding survey is the biographical record of Tiedemann, which appeared in 1787. Tiedemann deserves to be credited with a very real scientific priority. His record of his son's behavioral development was thoroughly scientific in spirit. Although Héroard's record precedes Tiedemann's, Héroard's interest was that of a diarist and historian; his account of the behavior of young Louis XIII was not oriented toward the study of human development.

Following Tiedemann, there seems to be a gap of forty-one years during which we have no record of any publication relating to normal child behavior. In 1828, Niederer published part of Pestalozzi's notes of a father. Thereafter, there was a consistent increase in publication. Between 1830 and 1860 there were seven publications. The 1860's saw the appearance of four titles. A sharp increase in publication occurred during the 1870's, the total number during that decade being seventeen. During the years 1880 and 1881 alone, 12 titles are listed in our bibliography.

The decade by decade comparison of publication indicates that an increase in activity toward the development of modern child psychology had been under way for some time before the publications of Preyer and Hall. These two scientists were apparently responding to the same influences of the times as were lesser men, but since Preyer and Hall were abler workers their contributions were more outstanding.

In this survey we have seen no cause to question the view that Preyer's *The Mind of the Child* and Hall's *Contents of Children's Minds* were the most notable contributions appearing before 1885. It has been suggested, however, that Galton's *Inquiry into the Human Faculty and Its Development* should be rated as almost equal to the afore-mentioned publications of Preyer and Hall in regard to its contributions to child psychology.

BIBLIOGRAPHY

1. ALCOTT, B., & PEABODY, ELIZABETH P. *Record of a school.* Boston, N. Y., Phila.: James Munroe & Co., 1835.
2. ALCOTT, B., & PEABODY, ELIZABETH P. *Conversations with children on the Gospels.* Boston: James Munroe & Co., Vol. 1, 1836, Vol. 2, 1837.
3. BARTHOLOMÄI, F. The contents of children's minds on entering school at the age of 6 years. *Städtisches Jahrbuch, Berlin und Seine Entwicklung*, 1870, **4**. (Translation published in Rep. U. S. Com. Educ. for year 1900–1901, **1**, 709–729. Washington: U. S. Government Printing Office, 1902.)
4. BIEDERT, P. Zur Technik der Pneumatometrie. *Dtsch. arch. f. klin. med.* 1875, **18**, 115–117.
5. CHAMPNEYS, F. H. Notes on an infant. *Mind*, 1881, **6**, 104–107.
6. CLARKE, E. H. *Visions: a study of false sight.* Boston: Houghton, Osgood, & Co., 1878.
7. CRUMP, L. *Nursery life 300 years ago.* London: Geo. Routledge & Sons. 1929.
8. CUIGNET, L. De la vision chez le tout jeune enfant. *Ann. Ocul.*, 1871, **66**, 117–126.
9. DARWIN, C. *The expression of the emotions in man and animals.* New York: Appleton, 1873.
10. DARWIN, C. Biographical sketch of an infant. *Mind*, 1877, **2**, 285–294.
11. DONDERS, F. C. Versuch einer genetischen Erklärung der Augen-Bewegungen. *Pflüg. Arch. f. d. ges. Physiol.*, 1876, **13**, 373–421.
12. EGGER, M. E. *Observations et reflexions sur le developpement de l'intelligence et du langage chez les enfants.* Paris: Picard, 1870.
13. FELDMAN, HENRICUS. *De statu normali functionum corporis humani animadversiones quandam.* Bonne: C. Georgie, 1833.
14. GALTON, F. Twins, as a criterion of the relative power of nature and nurture. *J. anthrop. Inst.*, 1875, **5**, 324–329.
15. GALTON, F. Psychometric experiments. *Brain*, 1879, **2**, 149–157.
16. GALTON, F. Statistics of mental imagery. *Mind*, 1880, **5**, 301–318.
17. GALTON, F. Visualized numerals. *J. Roy. anthrop. Inst.*, 1880, **10**, 85–102.
18. GALTON, F. *Inquiries into the human faculty and its development.* London: Macmillan, 1883.
19. GENZMER, A. *Untersuchungen über die Sinneswahrnehmungen des neugeborenen Menschen.* Halle: Max Niemeyer, 1882.
20. HALL, G. S. Contents of children's minds. *Princeton Rev.*, 1883, **11**, 249–272.
21. HÉROARD, JEAN. Journal, 1601–1628. Ed. by E. Soulie and E. de Barthelemy, 1868.
22. HERTZ, M. Einige Bermerkungen über das Saugen der Kinder. *Jahrb. f. Kinderhk.*, 1865, **7**, 46–57.
23. HICKS, J. B. On recording the fetal movements by means of a gastrograph. *Trans. obstet. Soc. London*, 1880, **22**, 134.
24. HOLDEN, E. S. On the vocabularies of children under two years of age. *Trans. Am. philol. Assoc.*, 1877, 58–68.
25. HUMPHREYS, M. W. A contribution to infantile linguistics. *Trans. Am. philol. Ass.*, 1880, **11**, 5–17. (Hartford: Case, Lockwood & Brainard Co., 1880.)
26. HUN, E. R. Singular development of language in a child. *Quart. J. psychol. Med.*, 1868, **2**, 525–528.
27. KRONER, T. Über die Sinnesempfindungen der Neugeborenen. *Breslauer aerzt. Zschl.*, 1882, **4**, 37–41.
28. KUSSMAUL, A. *Untersuchungen über das Seelenleben des neugeborenen Menschen.* Heidelberg & Leipzig: C. F. Winter, 1859.

29. LANGE, K. Der Vorstellungskreis unserer sechsjahrigen Kleinen. *Allgemeine Schultzeitung*, 1879, **56**, 327ff.

30. McCUSKEY, D. *Bronson Alcott, teacher*. New York: Macmillan, 1940.

31. MOLDENHAUER, W. Zur Physiologie des hörogans neugeborener. *Beitrage zur Geburtshülfe, Gynäkologie und Pädriatrik, Leipzig.* 1881, 199–204.

32. NIEDERER, JOHANNES. *Notes on Pestalozzi*. Aix-la-Chapelle, 1828. (Cited by R. De Guimps, *Pestalozzi, his life and work*. New York: Appleton, 1906. Pp. 40–48.)

33. POLLOCK, F. An infant's progress in language. *Mind*, 1878, **3**, 392–401.

34. POLLOCK, F. Review of Schulte's 'Die Sprache des Kindes.' *Mind*, 1881, **6**, 436–438.

35. PREYER, WILHELM. Psychogenesis. *Deutsche Rundschau*, 1880, **23**, 198–221. (Translated by Marion Talbot, *J. spec. Philos.*, 1881, **15**, 159–188.)

36. PREYER, W. *Die Seele des Kindes. Beobachtungen über die geistige Entwicklung des Menschen in den ersten Lebensjahren*. Leipzig: T. Grieben, 1882.

37. QUETELET, L. A. J. *Sur l'homme et le developpement des ses facultés, ou essai de physique sociale.* Paris: Bachdier, 1835.

38. RAELHMANN, E., & WITKOWSKI, L. Über atypische Augenbewegungen. *Arch. f. Anat. u. Physiol.*, 1877, **1**, 454–471.

39. SCHLEICHER, A. Einige Beobachtungen an Kindern. *Beitrage zur vergl. Sprachforsh.*, 1861, **2**, 497–498.

40. SCHULTZE, F. *Die Sprache des Kindes. Eine Anregung zur Erforschung des Gegenstandes.* Leipzig: Günther, 1880.

41. SIGISMUND, BERTHOLD. *Kind und Welt. Vatern, Muttern und Kindefrunden gewidmet.* Braunschweig: Vieweg, 1856.

42. SIKORSKY, A. J. Sur les effets de la lassitude provoquée par les travaux intellectuels chez les enfants de l'âge scolaire. *Ann. d'hygiene publ. Paris*, 1879, **2**, series 3, 458–464.

43. STRUMPELL, L. Notizen über die geistige Entwicklung eines weiblichen Kindes Während der ersten zwei Lebenjahr. In *Psychologische Pedagogik*, Leipzig: Georg Bohme, 1880.

44. SULLY, J. Babies and science. *Cornhill Mag.*, 1881, **43**, 539–554.

45. TAINE, H. Note sur l'acquisition de langage chez les enfants et dans l'espèce humaine. *Rev. phil.*, 1876, **1**, 3–23. (Translated in *Mind*, 1877, **2**, 252–257.)

46. TALBOT, E. *Papers on infant development.* Educ. Dept., Amer. Soc. Sci. Assoc., 1882, pp. 52.

47. TIEDEMANN, D. Beobachtungen über die Entwicklung der Seelenfahrigkeiten bei Kindern. *Hessichen Beitrage zur Gelehrsamkeit und Kunst*, 1787, **2**, 313 ff.; **3**, 486 ff. (Translated by S. Langer and C. Murchison, *J. genet. Psychol.*, 1927, **34**, 205–230.)

48. VIERORDT, K. *Der Gang des Menschen.* Tubingen: Laupp, 1881.

49. VIERORDT, K. Die Sprache des kindes. *Dtsche. Rev.*, 1879, **4**, 29–46.

50. WILLARD, MRS. EMMA. Observations upon an infant during its first year. By a mother. In *Progressive Education*, by Madame Necker de Saussure. Boston: Wm. D. Ticknor, 1835, pp. 323–348.

51. WYMA, ———. The mental development of the infant of today. *J. psychol. Med. & mental Pathol. New series*, 1881, **7**, 62–69.

Received March 3, 1949.

WIENER, NORBERT. *Cybernetics (or control and communication in the animal and the machine).* New York: John Wiley, 1948. Pp. ii+194.

This curious and provocative book is actually a series of essays on a variety of topics rather than an integrated work. It treats of such diversities as servo mechanisms, statistical mechanics, time, digital and analogue computers, physiology of the nervous system, perception, psychopathology, prosthesis, and the nature of society. On each of these seemingly unrelated subjects Dr. Wiener has something challenging to say. These concepts and subject matters represent to the author sub-topics in the new discipline of cybernetics. This term, which is formed from the Greek word meaning *steersman*, is defined as embracing the entire field of control and communication theory as applied to both animals and machines.

The author is professor of mathematics at the Massachusetts Institute of Technology. It is not surprising, therefore, to find that approximately one third of the book treats the sophisticated mathematics for which he is famous. In this part of the work, which will probably be fully understood only by professional mathematicians, the author develops the communication theory which he proposes as a basic theoretical framework for treating the activity of complex machines and living organisms. Beginning this portion with a brief history of the statistical mechanics of Gibbs and Lebesgue, Dr. Wiener discusses the development of ergodic theory under Koopman and von Neumann, then speaks briefly of entropy and the Maxwell demon and finally passes on to the development of a statistical mechanics of time series. It is the latter which the author regards as most appropriate in discussing animal behavior.

Fortunately for the majority of readers, mathematical notation is held to a minimum in the remainder of the book. Sometimes with great simplicity, but always brilliantly, the author hurries from discussions of the concept of time, through rather technical yet fascinating discussions of servo mechanisms and electronic computers, through essays on psychopathology and human perception to pessimistic prognostications concerning the fate of man and his society. One is a little breathless when he arrives at the end of the last chapter where, incidentally, he discovers a final note on how to construct a chess playing machine.

There are many challenging speculations in this work which will provoke discussions and, it is to be hoped, research. But the author's most important assertion for psychology is his suggestion, often repeated, that computers, servos, and other machines may profitably be used as models of human and animal behavior. Although it is nowhere stated explicitly, the implication is strong that the value to the physiologist and psychologist of these physical models lies less within them-

selves than in the ready-made mathematical theory which is now associated with them. It seems that Dr. Wiener is suggesting that psychologists should borrow the theory and mathematics worked out for machines and apply them to the behavior of men.

Action upon such a proposal is, of course, contingent upon evidence of the degree of similarity between the analogue and that to which it is supposed to be analogous. A close correspondence between the two would suggest that there might be some profit in applying the mathematics of the one to the other, though no degree of correspondence short of perfect identity could relieve the researcher of his scientific caution.

Obviously, since men can not as yet fabricate themselves out of vacuum tubes and circuits there is less than perfect identity between man and the models proposed by Dr. Wiener. But is the parallel sufficiently close to be encouraging?

Little evidence on this point is cited in the book other than examples of gross similarity between the action of men and machines. Indication is given of a parallel between voluntary activity and the response of a closed loop servo which responds, not to an absolute signal, but to the difference between a desired state of affairs and the state of affairs which exists at the moment. Certain forms of ataxia are recognized as suggesting certain conditions of servo instability, while others are said to resemble the response of a servo with an open feedback loop. The activity of the brain is compared to the working of an electronic computer and the suggestion is made that certain forms of human insanity have their parallel in computer breakdowns.

These and other like observations constitute almost the entire body of "evidence" that machines closely resemble men in their activities. Almost no factual data other than of a purely physiological nature exist to lend credence to this important postulate.

In fairness to the author it must be pointed out that no attempt is made to "prove" the fruitfulness of the proposed analogous reasoning nor to provide an exhaustive catalogue of man-machine correspondences. Dr. Wiener is content to develop the mathematical theory of machines and then to point out certain similarities between the responses of these devices and those of living beings. He may regard it as the task of others to fill in the details of the pattern which he has sketched in outline.

Yet one cannot help but regret that more of the details were not at hand when the book was written. In the final analysis it is precisely these particulars which will determine whether the book should be judged as literature or science. At present one can only hope that cybernetics will prove to be as fruitful as it now stimulating to the imagination.

Franklin V. Taylor.

Naval Research Laboratory.

BOOKS AND MATERIALS RECEIVED

ANDREWS, T. G., *Some psychological apparatus: A classified bibliography.* Psychological Monographs, No. 289. Washington, D. C.: American Psychological Assn.: 1948. Pp. 38.

ANSBACHER, H. L. *Attitudes of German prisoners of war: a study of the dynamics of national-socialistic followership.* Psychological Monographs, No. 288. Washington, D. C.: American Psychological Assn. 1948. Pp. 42. $1.00

BLUM, MILTON L. *Industrial psychology.* New York: Harper, 1949. Pp. xi+519. $4.50.

BOUTHILET, LORRAINE, AND BYRNE, KATHERINE MANN. *You and your mental abilities.* Chicago: Science Research Associates, 1948. Pp. 48.

BRECKENRIDGE, MARIAN E., AND VINCENT, E. LEE. *Child development.* Philadelphia: W. B. Saunders, 1949. Pp. 622. $4.00.

BROUWER, PAUL J. *Student personnel services in general education.* Washington, D. C.: American Council on Education, 1949. Pp. xix+317. $3.50.

BUROS, OSCAR K. *The third mental measurements year book.* New Brunswick: Rutgers Univ. Press, 1949. Pp. xiv+1246. $12.50.

DAS, SONYA RUTH. *The American woman in modern marriage.* New York: Philosophical Library, 1948. Pp. 185. $3.75.

DE FORD, MIRIAM ALLEN. *Psychologist unretired.* Stanford Univ., Calif.: Stanford Univ. Press, 1948. Pp. ix+127. $3.[illegible]

DE GRAZIA, SEBASTIAN. *The political commun[illegible].* Chicago: Univ. of Chicago Press, 1948. Pp. xx+258. $4.00.

ERICKSON, CLIFFORD E. *A practical handbook for school counselors.* New York: Ronald Press, 1949. Pp. vii+224. $3.00.

ESCALONA, SIBYLLE K. *An application of the level of aspiration experiment to the study of personality.* New York: Bureau of Publications, Teachers College, Columbia Univ., 1948. Pp. viii+132. $2.10.

HARING, DOUGLAS G. *Personal character and cultural milieu.* (Rev. Ed.) New York: Syracuse Univ. Press, 1949. Pp. x+670. $5.00.

HARRIS, NOEL G. *Modern trends in psychological medicine.* London: Paul B. Hoeber, 1948. Pp. xii+450. $10.00.

HAVIGHURST, ROBERT J., AND TABA, HILDA. *Adolescent character and personality.* New York: John Wiley, 1949. Pp. x+315. $4.00.

HOLLINGWORTH, HARRY L. *Psychology and ethics.* New York: Ronald Press, 1949. Pp. ix+247. $3.50.

HUSBAND, RICHARD W. *Applied psychology.* New York: Harper, 1949. Pp. xiii+845.

JENNINGS, HELEN HALL. *Sociometry in group relations.* Washington, D. C.: American Council on Education, 1948. Pp. ix+85. $1.25.

KANNER, LEO. *Child psychiatry.* (2nd Ed.) Springfield, Ill.: Charles C Thomas, 1949. Pp. xxiv+752. $8.50.

KLEIN, MELANIE. *Contributions to psychoanalysis.* London: Hogarth Press, 1948. Pp. 416.

KUNST, MARY S. *A study of thumb- and finger-sucking in infants.* Psychological Monographs, No. 290. Washington. D. C.: American Psychological Assn., 1948.

LENNEP, D. J. VAN. *Four picture test.* The Hague: Martinus Nijhoff, 1948. Pp. 48.

MCNEMAR, QUINN. *Psychological statistics.* New York: John Wiley, 1949. Pp. vii+364. $4.50.

MACHOVER, KAREN. *Personality projection.* Springfield, Ill.: Charles C Thomas, 1949. Pp. ix+191. $3.50.

MENNINGER, WILLIAM C. *Understanding yourself.* Chicago: Science Research Associates, 1948. Pp. 52.

MINER, ROY WALDO, *et al. Newer synthetic analgesics.* New York: New York Academy of Sciences, 1948. Pp. 174.

MURPHY, GARDNER. *Historical introduction to modern psychology.* (Rev. Ed.) New York: Harcourt, Brace, 1949. Pp. xiv+466.

ROTMAN, DAVID B. *Addresses and papers.* Chicago: Psychiatric Institute, 1948. Pp. xiv+204.

SADLER, WILLIAM S. *Adolescence problems.* St. Louis: C. V. Mosby, 1948. Pp. 466. $4.75.

SCHILLER, PAUL V. *Aufgabe der psychologie.* Wein: Springer Verlag, 1948. Pp. 233.

SEGEL, DAVID. *Intellectual abilities in the adolescent period.* Washington: Federal Security Agency (Office of Education), Bulletin 1948, No. 6.

SMITH, GUDMUND. *Psychological studies in twin differences.* Lund, Sweden: C. W. K. Gleerup, 1949. Pp. 251.

SWINEFORD, FRANCES. *A study in factor analysis.* Chicago: Univ. of Chicago Press, 1948. Pp. xi+70. $2.00.

SYMONDS, PERCIVAL M. *Symond's picture-story test.* New York: Bureau of Publications, Teachers College, Columbia Univ. 1948. Pp. 20. $5.25.

SYMONDS, PERCIVAL M. *Dynamic psychology.* New York: Appleton-Century-Crofts, 1949. Pp. vii+413. $3.75.

REPORT TO LIBRARIANS

Available Back Issues of the Psychological Bulletin

Thirty-three issues of the Psychological Bulletin became out-of-print during 1948. Libraries and institutions ordering "All Available Issues" ordered 10 sets of this journal.

Estimating a similar number of sales of all available issues during 1949, the inventory shows that 17 other issues will become out-of-print during the next year; that is, there are as of January 2, 1949, 17 issues of fewer than ten copies each.

Copies of this journal are stored in the warehouses of the two printers who have printed it between 1904 and 1949. About 390 separate issues are handled for an "All Available Issues" sale. Three weeks should be allowed as a minimum before expecting such an order to be filled. For quicker handling, the APA keeps issues of the last two years available in the APA office.

American Psychological Association
1515 Massachusetts Avenue Northwest
Washington 5, D.C.

S

rience,
r, the
, psy-
ful to
other
$2.75

func-
close
imag-
magi-
f con-
fe and
e ex-
$3.75

e in
osed